INFORME FINAL

PROYECTO INTERNO

COORDINACIÓN DE HIDROLOGÍA
SUBCOORDINACIÓN DE HIDROMETEOROLOGÍA

Desarrollo de una herramienta de pronóstico estacional con fines agrícolas para la precipitación y el escurrimiento: Estudio piloto en una cuenca del Noroeste de México.

Elaboró:

Martín José Montero Martínez
Waldo Ojeda Bustamante
Iván Rivas Acosta
Julio Sergio Santana Sepúlveda
Patricia Guadalupe Herrera Ascencio

Con la colaboración de:

Floris van Ogtrop (U. Sydney, Australia)
Willem Vervoort (U. Sydney, Australia)

17 de diciembre 2014

Contenido

1. Resumen Ejecutivo

Este documento presenta los resultados del esfuerzo conjunto de un grupo de investigadores del IMTA con otro de la Universidad de Sídney para el desarrollo de una herramienta de pronóstico estacional estadístico de escurrimiento y precipitación en la cuenca del Río Huites en el Noroeste de México. La herramienta desarrollada tiene potencial de aplicación para los sectores hídrico y agrícola en donde esta información resulta muy valiosa para llevar a cabo una mejor planeación tanto de la disponibilidad del recurso hídrico como la planeación de un año agrícola para una región determinada.

Los métodos estadísticos aquí empleados para la generación de la herramienta están basados en los modelos estadísticos generalizados, tanto el escalar (GAM) como el vectorial (VGAM) los cuales suelen ser métodos de ajuste de mayor complejidad que los típicos métodos de regresión lineal múltiple o los modelos lineales generalizados. Los GAM son modelos de regresión similares a los GLM pero en los que la relación de variable dependiente con las variables independientes sigue una distribución de probabilidad de la familia exponencial y además toma en cuenta de mejor forma la contribución "aditiva" de otras variables de valor importante (como pueden ser las temperaturas de la superficie del mar de las diferentes regiones de El Niño o los índices de oscilaciones climáticas) para determinar la variable dependiente, que en nuestro caso será la precipitación y el escurrimiento.

Por otro lado, se escogió la Cuenca del Río Huites como la zona de estudio debido a que este tipo de cuenca tiene poca influencia antropogénica en la determinación del producto integral de escurrimiento por influencia directa de las precipitaciones registradas en la cuenca. Esto permite que los resultados del modelo estadístico tengan una influencia más directa entre los potenciales cambios de las variables climáticas y su influencia en la precipitación total y escurrimiento registrado en la parte más baja de la cuenca.

Los análisis de resultados del presente estudio arrojan como mejor modelo de pronóstico del escurrimiento máximo (en Septiembre) de la cuenca con la combinación de un modelo GAM con el flujo, la precipitación, las temperaturas de superficie del mar (TSM) de la zona Niño 1+2 y el índice de Oscilación del Ártico de dos meses de anticipación (esto es del mes de Julio anterior) para el pronóstico continuo. Para el pronóstico categorizado el mejor modelo resulto ser la misma combinación anterior excepto de cambiar las TSM de El Niño 1+2 por las TSM de El Niño 4. Estos resultados fueron corroborados por estrictas métricas que evalúan la calidad del pronóstico como el error medio absoluto y la eficiencia Nash-Sutcliffe para el pronóstico continuo; así como las tablas de contingencia y el índice de habilidad de Heidke para el pronóstico categorizado.

En cuanto al pronóstico por terciles del escurrimiento total acumulado de Junio–Septiembre y de Septiembre–Diciembre, períodos de relevancia agrícola para la zona de estudio, el mejor modelo se obtuvo de la aplicación del modelo VGAM con la señal de Niño 1+2, IOD y AO promediada de Febrero–Abril para el primero (Junio–Septiembre), y una combinación de la señal del Flujo, Niño 1+2, IOD y AMO promediada de Mayo–Julio para el segundo (Septiembre–Diciembre).

En el futuro próximo esperamos aplicar esta metodología para otras cuencas relevantes en México.

2. Introducción

El Servicio Meteorológico Nacional (SMN) actualiza y publica mensualmente el pronóstico estacional oficial para México con un horizonte para los próximos tres meses. Actualmente, es el resultado de la integración y discusión de especialistas del SMN, mediante el uso de métodos de pronósticos estadísticos ("Años Análogos" y la herramienta "Climate Predictability Tool", CPT) y pronósticos dinámicos (Modelos numéricos del clima de diversos Centros Internacionales y ensambles multimodelo).

El método de "Años Análogos" identifica patrones oceánicos y atmosféricos de años en el pasado los cuales tienen similitud con las condiciones actuales. El CPT aplica la "Regresión de Componentes Principales" para identificar los patrones más representativos en el océano y correlacionarlos con la precipitación en México. Los modelos dinámicos considerados por los especialistas del SMN corresponden a la simulación numérica de la atmósfera y el océano (CFSv2, NASA-GMAO, NCAR, etc.), cuyos resultados se obtienen de Centros Internacionales del Clima. La perspectiva estacional del clima se actualiza en los primeros días de cada mes.

Sin duda un buen pronóstico estacional del caudal sería de mucha ayuda en salvar vidas y propiedades en México. En general, el pronóstico de caudales proporciona muchos beneficios a la sociedad, mediante la mejora de nuestra capacidad de planificar y adaptarse a los cambios del suministro de agua. Un enfoque común para el desarrollo de estas previsiones es el uso de técnicas estadísticas que relacionan un conjunto de predictores que representan el estado del clima en su relación con el caudal histórico y, a continuación, utilizando este modelo para proyectar los caudales de una o más temporadas de antelación sobre la base de la corriente o de un estado climático proyectado (Westra et al., 2008).

Estudios revisados por pares sobre la relación entre temperatura de superficie del mar - lluvia - escorrentía en México no son muchos. Sin embargo, hay algunos trabajos interesantes ya realizados en zonas específicas. Gochis et al. (2007) exploran la compleja relación entre las precipitaciones y los caudales en la región del Monzón de América del Norte (NAM) mediante la construcción de una hidroclimatología regional de cuencas de cabecera seleccionadas en el Noroeste de México. El régimen de precipitaciones NAM presenta una considerable variabilidad en escalas de tiempo interanuales a interdecadales que están potencialmente vinculadas a través de teleconexiones al forzamiento remoto. Temperaturas de la superficie del mar (SST) en el Pacífico norte (por ejemplo, Higgins y Shi, 2001; Englehart y Douglas, 2002;. Brito-Castillo et al, 2002) y en el Golfo de California (Mo y Juang, 2003), cada uno se correlacionan poco con precipitaciones NAM.

Mediante el análisis de correlación de teleconexión, Englehart y Douglas (2002) mostraron que la región oeste de la Sierra Madre Occidental (SMO) se correlaciona moderadamente con El Niño / Oscilación del Sur (ENOS), pero sólo durante la fase positiva de la Oscilación Decadal del Pacífico (PDO), lo cual no es el caso para el altiplano o meseta del centro de México. La correlación significativa entre el ENOS y la precipitación durante la fase negativa de la PDO no se encuentran en

ninguna de las regiones. Lo anterior apoya firmemente el argumento de que es importante incluir índices PDO en nuestro análisis.

La CONAGUA tiene un esquema convencional para acopiar la información estadística de producción agrícola de los distritos de riego del país basado principalmente en compilación de información en campo al finalizar los ciclos agrícolas. En las unidades de riego del país, la estadística agrícola es obtenida por diferencia de la publicada por la SAGARPA y la publicada por CONAGUA, para la agricultura de riego y para los distritos de riego, respectivamente.

El pronóstico estacional de la precipitación y temperatura permite realizar el pronóstico de cambios en el desarrollo de los cultivos y de las variables agrícolas asociadas a su producción. Por lo cual el desarrollo de una metodología de pronóstico estación es importante por varias razones:

- Para toma de decisiones anticipadas para optimizar insumos y maximizar el rendimiento de los cultivos.
- Para la planeación de gran escala a nivel nacional y regional. El pronóstico del rendimiento facilita el manejo de asuntos de seguridad alimentaria.
- El desarrollo de los cultivos se asocia a la demanda hídrica de los cultivos. Un mejor monitoreo del desarrollo de los cultivos permite una mejor gestión del agua en las zonas bajo riego.
- La precipitación está asociada a los escurrimientos potenciales de una cuenca, que son lo que se almacenan en las presas y sirven para suministrar el riego a los cultivos durante la temporada seca, por lo que el pronóstico estacional de caudales permitirá elaborar planes de riegos con mayor robustez.

La perspectiva del pronóstico estacional ciertamente puede contribuir a aportar en reducir la incertidumbre en el clima y que los cálculos agroclimáticos en lugar de trabajar solo con la información de las normales climáticas puedan comenzar a trabajar con la anomalía esperada durante los siguientes meses.

Esto significa que se necesita (ya sea espacial o temporalmente) un cierto nivel de escalamiento. Este problema no es nuevo, y es importante no sólo para la aplicación sugerida, sino también en el contexto de los grandes modelos de circulación global de escala y los estudios sobre el cambio climático. En este sentido el IMTA requiere de la cooperación internacional para generar productos con mayor rapidez y respondan a la necesidad de estimar la producción agrícola de las zonas de riego del país con oportunidad y calidad.

- *Colaboración con la Universidad de Sídney*

Investigadores académicos de la Universidad de Sídney han estado trabajando activamente en el desarrollo de modelos estadísticos para la predicción y previsión. En un esquema de cooperación bilateral en la modalidad de proyecto, financiado por AusAID recientemente concluido, el personal trabajó con investigadores del IMTA en esta área para ampliar el conocimiento.

Para este proyecto, los investigadores del IMTA recopilarán los datos y construirán los modelos estadísticos con la orientación de los investigadores de la U. de Sídney. Toda la comunicación será a través de conexión remota.

2.1 Objetivos

Explorar la posibilidad de encontrar una herramienta útil para predecir el comportamiento de la escorrentía estacional para periodos de 6 meses a 1 año de antelación, que permita mejorar la planificación de un año agrícola en zonas de riego.

Detectar correlaciones significativas entre SST - Precipitación y SST - Escorrentía para varias estaciones climatológicas e hidrométricas en el Noroeste de México.

3. Metodología

3.1 Consulta de fuentes de información y recopilación de datos

Para la elaboración de este informe se revisó información de fuentes diversas, como bases de datos, informes, publicaciones científicas, catálogos, páginas de internet con documentación oficial y talleres asociados al tema de cambio climático, así como proyectos y programas nacionales e internacionales asociados al cambio climático.

3.1.1 Caracterización de la variabilidad climática en México

El clima es la síntesis de la temperie en una región particular, éste puede ser definido cuantitativamente usando los valores esperados de los elementos meteorológicos en un sitio durante un mes o un periodo mayor; dichos valores esperados pueden ser llamados elementos climáticos e incluyen variables como la temperatura promedio, precipitación, viento, presión, nubosidad y humedad. En la definición del clima se emplean los valores de estos elementos en superficie y de estos, los de mayor interés meteorológico son la temperatura y la precipitación (Hartmann, 1994).

- Controles permanentes del clima

De acuerdo con Mosiño y García (1974), los controles permanentes del clima son aquellas características físicas propias de un sitio que influyen en los fenómenos meteorológicos; como: latitud, orografía y su orientación y relación respecto a cuerpos de agua principalmente océanos (Mosiño y García, 1974; Hartmann, 1994).

La importancia de la latitud es debida a la variación de la insolación a lo largo del año (Mosiño y García, 1974). La temperatura superficial es más grande cerca del Ecuador, donde excede los 26°C a lo largo de una ancha banda de latitudes; fuera de este cinturón la temperatura superficial decrece de forma constante hacia ambos polos. En el Hemisferio Norte (HN) se observa una fuerte variación estacional de la temperatura. La amplitud del ciclo estacional decrece del polo hacia el Ecuador donde la temperatura media zonal permanece alrededor de 27°C (Hartmann, 1994).

Los principales efectos de la orografía en la atmósfera en México se resumen en: efectos de represa, efectos de desviación, efectos de bloqueo, altitud sobre el nivel del mar, ascenso forzado y calentamiento adiabático por descenso; estos efectos no son independientes unos de otros (Mosiño y García, 1974). Precipitaciones intensas y persistentes pueden resultar cuando el aire húmedo es forzado a subir por cordilleras montañosas por vientos persistentes (Hartmann, 1994).

- Patrones dominantes de variabilidad climática

Los fenómenos atmosféricos característicos se ven modulados por mecanismos que tienen periodicidades desde menos de una estación (intraestacionales), de una estación a otra

(interestacionales), de menos de un año (intraanuales), de algunos años hasta menos de una década (interanuales), de algunas décadas hasta menos de un siglo (decadales) y así sucesivamente hasta la escala glacial-interglacial.

- Ciclo diurno y anual

Los ciclos diurno y anual de las variables atmosféricas se deben a cambios en la radiación recibida en la superficie como resultado de los movimientos de la Tierra. La rotación de la Tierra sobre su eje origina el ciclo diurno y el movimiento de traslación de la Tierra alrededor del Sol, el ciclo anual; el elemento más importante en la producción del ciclo anual es la cantidad de radiación recibida en un sitio a través del año la cual varía dependiendo de la intensidad y duración de la luz solar en el sitio. La intensidad es principalmente función del ángulo de incidencia que afecta la energía recibida por unidad de área y la cantidad de energía absorbida. La intensidad de la radiación, es máxima en latitudes donde la radiación solar es perpendicular a la superficie (cenit). El cenit se mueve hacia el norte y el sur a través del año hasta 23°30'; los trópicos son las latitudes más lejanas al Ecuador donde la radiación es perpendicular al menos una vez al año (Oliver y Hidore, 2002).

- El Niño – Oscilación del Sur

El Niño es una fuente dominante de variabilidad climática interanual alrededor del mundo; es una condición anómala de la temperatura del océano en el Pacífico tropical del este, la componente atmosférica se denomina Oscilación del Sur, por lo que a menudo el fenómeno es llamado El Niño-Oscilación del Sur (ENSO), haciendo referencia al proceso acoplado océano-atmósfera; El Niño entonces corresponde a la fase cálida del ENSO y La Niña a la fase fría (Trenberth, 1997) sin que sea necesaria una secuencia de fases frías y cálidas (Vázquez-Aguirre, 2007).

El Niño se cuantifica en términos de índices que corresponden a anomalías, cuando la temperatura superficial del mar (SST por sus siglas en ingles) en la región Niño-3 (5°N-5°S, 150°-90°W) excede en 0.5°C o cuando la anomalía de la SST en la región Niño 3-4 (5°N-5°S, 170°-120°W) excede 0.4°; esto es suficiente para producir impactos en los países de la costa del Pacífico (Trenberth y Stepaniak, 2001) (**Fig. 3.1**). Esta anomalía positiva de la temperatura superficial del mar en el Pacífico ecuatorial del este, altera la localización e intensidad de las regiones de convección profunda en los trópicos y sus efectos en la circulación atmosférica global. El incremento en la temperatura en las aguas superficiales es parte de la respuesta oceánica a las condiciones atmosféricas alteradas, especialmente los cambios en los vientos alisios sobre el Pacífico (Philander, 1989).

De acuerdo con Magaña (2004), de manera general, las lluvias de invierno se intensifican durante años El Niño en el noroeste y noreste de México, mientras que disminuyen hacia la parte sur (Magaña et al., 1998). Los inviernos de El Niño resultan más fríos en casi todo el país. Por otra parte, los veranos de El Niño son más secos y cálidos que los veranos de La Niña.

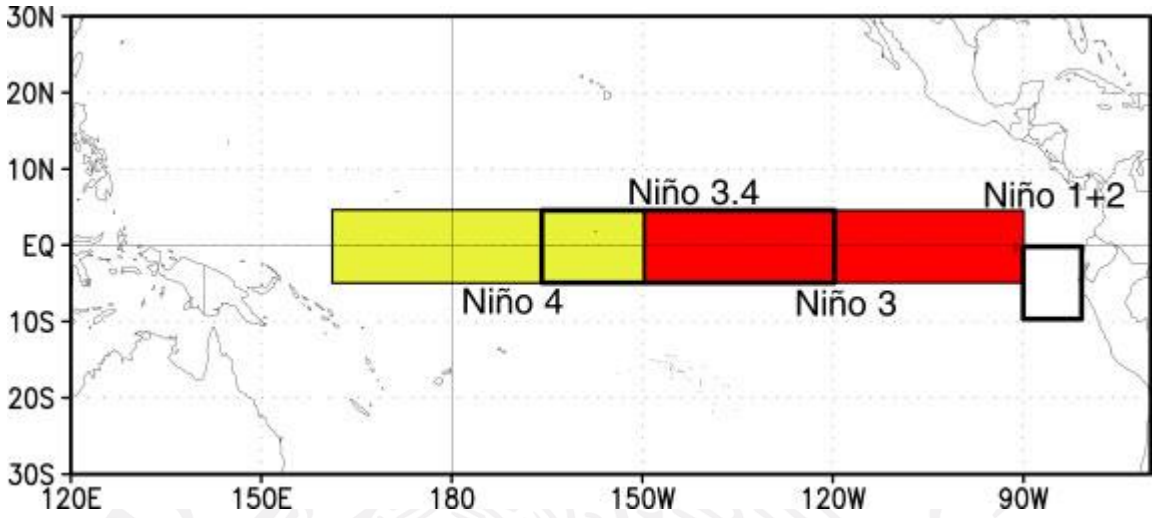


Fig. 3.1 Delimitación de las cuatro zonas principales de El Niño.

- Oscilación Decadal del Pacífico

De acuerdo con Mantua (2001), en el siglo XX, los eventos de la Oscilación Decadal del Pacífico (PDO) tuvieron una persistencia de 20 a 30 años y estos presentan su principal huella climática visible en el Pacífico Norte y el sector de Norteamérica y la segunda en los trópicos. Las fases cálidas de la PDO prevalecieron durante 1890-1924 y 1947-1976 y las fases cálidas durante 1925-1946 y de 1977 a mediados de los 90s (Mantua et al. 1997, Minobe 1997). Minobe (1999) mostró que durante el siglo XX las fluctuaciones de la PDO fueron principales en dos periodicidades generales, uno de 15 a 25 años y otra de 50 a 70.

El patrón en la fase cálida de la PDO es SST anómalamente fría en la parte central del Pacífico Norte coincidente con SSTs cálidas a lo largo de la costa oeste de América con anomalías de la presión al nivel del mar (PNM) que varían en un patrón de onda con bajas presiones sobre el Pacífico Norte y altas PNM sobre el oeste de Norteamérica y el Pacífico subtropical; estos patrones de presiones causan viento en el sentido contrario a las manecillas del reloj sobre el Pacífico Norte.

La combinación de información de PDO y ENSO puede aumentar la destreza para pronóstico climático de Norteamérica, dado que la influencia de ENSO en dicha región depende de la fase de PDO ya que cuando estos se encuentran en fase se presentan eventos extremos.

- Oscilación Multidecadal del Atlántico

La Oscilación Multidecadal del Atlántico (AMO) es un modo de variabilidad que ocurre en el Océano Atlántico Norte, que tiene su principal expresión en el campo de la temperatura de superficie del mar (TSM) (Dijkstra et al., 2006). Si bien existe cierto apoyo a esta modalidad en los modelos y en observaciones históricas, existe controversia en cuanto a su amplitud y, en particular, la atribución del

cambio de la temperatura de superficie del mar a causas naturales o antropogénicas, especialmente en las zonas tropicales del Atlántico importantes para el desarrollo de huracanes (Mingfang et al., 2009).

La AMO fue identificada por Schlesinger y Ramankutty en 1994.

La señal de AMO se define generalmente a partir de los patrones de variabilidad SST en el Atlántico Norte, una vez se ha eliminado cualquier tendencia lineal. Este *detrending* pretende eliminar la influencia del calentamiento global inducido por los gases de efecto invernadero a partir del análisis. Sin embargo, si la señal de calentamiento global es significativamente no lineal en el tiempo (es decir, no sólo un aumento lineal suave), las variaciones en la señal forzada se filtran a la definición AMO. En consecuencia, las correlaciones con el índice AMO pueden enmascarar los efectos del calentamiento global.

El índice de AMO fue definido por Enfield et al. (2001), como la media móvil de 10 años de la anomalía de temperatura superficial del mar en el Atlántico al norte del Ecuador. Delworth y Mann (2000), demostraron que tiene picos espectrales de entre 50 y 70 años de frecuencia; de acuerdo con Enfield et al. (2001) las fases cálidas de AMO ocurrieron durante 1860-1880 y 1940-1960 y las fases frías durante 1905-1925 y 1970-1990.

- Oscilación del Atlántico Norte

La Oscilación del Atlántico Norte es una fluctuación a gran escala en la masa atmosférica situada entre la zona de altas presiones subtropicales y la baja polar en la cuenca del Atlántico Norte. Su influencia se extiende desde Norteamérica Central a Europa, alcanzando incluso al Norte de Asia.

Determina la variabilidad de clima invernal en la región del Atlántico Norte y se estima mediante el índice NAO, que se calcula como la diferencia de presión a nivel del mar que se produce entre las bajas presiones de Islandia y las altas presiones de la Azores. El nombre fue citado por primera vez por Walker G.T en 1924.

Aunque el índice NAO varía anualmente, también presenta una tendencia a quedarse en una fase (positiva o negativa) durante intervalos de varios años.

El índice NAO se ha estimado como diferencias de la presión a nivel del mar normalizadas entre Lisboa y el mar de Stykkisholmur para los valores medios del periodo de Diciembre a Marzo. Los valores se han normalizado respecto al periodo 1864-1983.

La fase positiva del índice NAO indica un centro de alta presión subtropical más fuerte de lo normal y una depresión polar más profunda de lo normal.

El incremento de la diferencia de presión entre ambos centros de acción da como resultado que una mayor cantidad de tormentas de invierno y más fuertes, crucen el Océano Atlántico siguiendo una dirección más hacia el norte de lo normal.

Como resultado en Europa del Norte los inviernos son más cálidos y lluviosos mientras que en Canadá del norte y Groenlandia los inviernos son más fríos y secos. Las zonas orientales de los EE.UU experimentan condiciones de invierno templadas y húmedas mientras que en el Sur de Europa son secas.

La fase negativa del índice NAO se produce cuando se debilita el centro de altas presiones subtropicales y la depresión polar.

La reducción del gradiente de presión da como resultado que sean menos las tormentas de invierno y más débiles las que crucen la cuenca del Atlántico Norte en dirección Este-Oeste, aportando aire húmedo a la cuenca del Mediterráneo y aire frío a Europa del Norte.

La costa este de los EE.UU experimenta condiciones de invierno con más irrupciones de aire frío y por lo tanto más temporales de nieve. Sin embargo, en Groenlandia las temperaturas invernales son más templadas.

- Oscilación del Ártico

La Oscilación del Ártico (AO) se refiere a un patrón opuesto de presión entre el Ártico y las latitudes medias del norte. En general, si la presión atmosférica es alta en el Ártico, tiende a ser baja (la AO) en las latitudes medias del norte, como el norte de Europa y América del Norte. Si la presión atmosférica es baja en las latitudes medias a menudo es alta en el Ártico. Cuando la presión es alta en el Ártico y baja en las latitudes medias, la AO se encuentra en su fase negativa. En la fase positiva, el patrón se invierte. (**Fig. 3.2**)

Los meteorólogos y climatólogos que estudian el Ártico prestan atención a la Oscilación del Ártico, ya que su fase tiene un efecto importante sobre el clima en lugares del norte. La fase positiva de la AO trae tormentas oceánicas más al norte, haciendo el tiempo más húmedo en Alaska, Escocia y Escandinavia y más seco en el oeste de Estados Unidos y el Mediterráneo. La fase positiva también mantiene un tiempo más cálido de lo normal en el este de Estados Unidos, pero más frío de lo normal en Groenlandia.

En la fase negativa de la AO los patrones se invierten. Una fase muy negativa de la AO trae un tiempo cálido hacia latitudes altas, y tiempo más frío y tormentoso a las regiones más templadas, donde vive la gente. Durante gran parte del siglo pasado, la AO alternó su fase positiva y negativa. Durante el período de la década de 1970 a mediados de 1990, la Oscilación del Ártico tendió a quedarse en su fase positiva. Sin embargo, desde entonces ha alternado de nuevo entre positivo y negativo, con una fase negativa de registro en el invierno de 2009-2010.

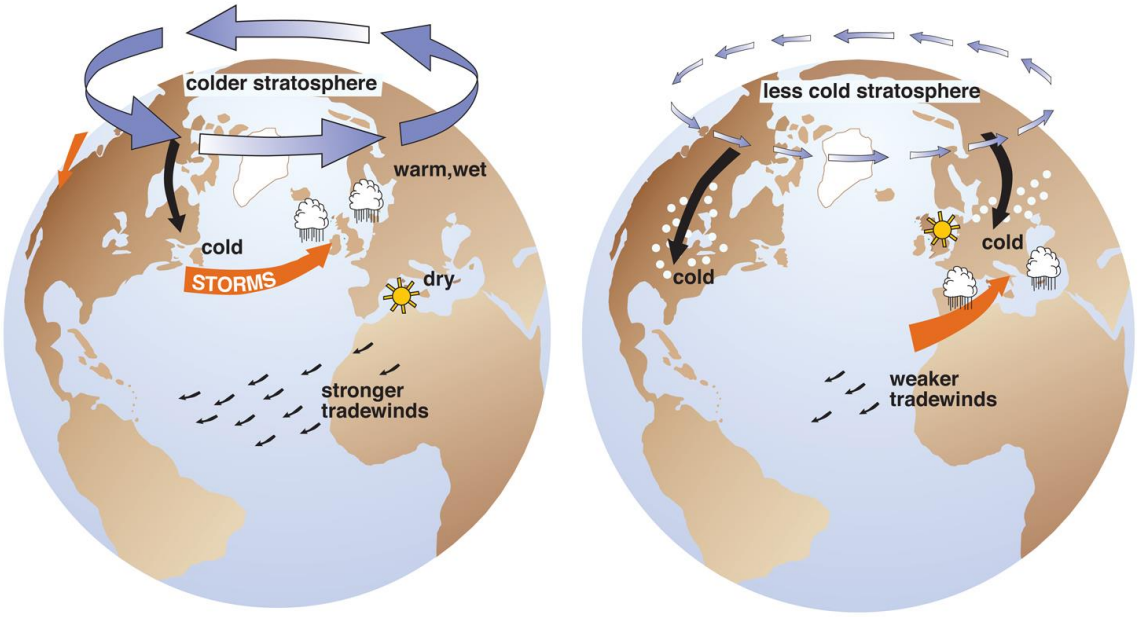


Fig. 3.2 Efectos de la Fase Positiva (izquierda) y Negativa (derecha) de la Oscilación del Ártico. Crédito: J. Wallace (U. of Washington).

3.1.2 Caracterización de la región de estudio – Cuenca del Río Huites

La cuenca del Rio Huites es una sub-cuenca del Rio Fuerte y se localizada en la Región Hidrológica No. 10, de acuerdo a la regionalización hidrológica de la Comisión Nacional del Agua. El Rio Fuerte es de gran importancia en esta región, tanto por su extensión como por los escurrimientos que en ella se generan. La cuenca está conformada por parte de los estados de Chihuahua, Durango, Sinaloa y Sonora, abarca una superficie total de 39,590 km^2. El área de drenaje de la presa Huites es de aproximadamente 26,047 km^2. (**Fig. 3.3**)

El Río Fuerte tiene una longitud de 540 kilómetros; presenta un orden máximo de 6 y un tipo de drenaje angulado. La cuenca es exorreica con una altura máxima de 3,198 metros sobre el nivel del mar. Esta porción de región hidrológica se localiza en los estados de Chihuahua, Durango, Sinaloa y Sonora; y está constituida por la corriente principal del mismo nombre y tiene un desarrollo a lo largo del colector general hasta la desembocadura al Golfo de California. Tiene su origen en un punto situado en el estado de Durango, que es común a los parteaguas de los Ríos Nazas y Culiacán.

En esta cuenca opera conjuntamente el sistema de presas Huites (oficialmente Luis Donaldo Colosio), Miguel Hidalgo y Josefa Ortiz de Domínguez, las cuales permiten el aprovechamiento integral del Rio Fuerte ya que regula las avenidas que se presentan por la influencia de deshielo o fusión de la nieve, lo cual se refleja en el incremento importante en los picos de avenidas de los hidrogramas, con gastos superiores hasta cuatro veces en los meses invernales o por el efecto de tormentas tropicales o ciclones durante el verano que conllevan grandes daños a los cultivos, a la infraestructura de riego, a las vías de comunicación, la ganadería e inclusive, a las poblaciones ribereñas.

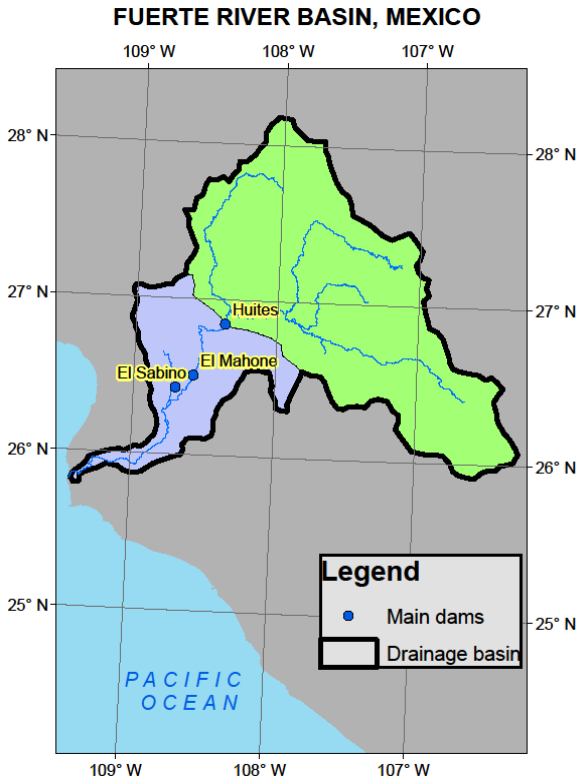


Fig. 3.3 Cuenca de la presa Huites

Por otro lado, mediante la operación conjunta de las presas Huites, Miguel Hidalgo y Josefa Ortiz de Domínguez permite abastecer las demandas de los Distritos de Riego No. 075 Valle del Fuerte y No. 076 Valle del Carrizo en el estado de Sinaloa. Además del beneficio al sector agrícola, las presas Huites y Miguel Hidalgo cuentan cada una con una central hidroeléctrica, aprovechando la carga del almacenamiento de las presas para la generación de energía eléctrica. En particular la presa Huites se ubica en Choix, Sinaloa y entró en operación el 15 de septiembre de 1996 y cuenta con una central hidroeléctrica capaz de generar 422 megawatts de energía eléctrica. Su embalse tiene un volumen aproximado de 2,908 hm^3.

Los distritos de riego cuentan con la infraestructura esencial para la gestión del agua para el riego y se resumen a continuación. *Distrito de Riego No. 76 Valle del Carrizo*, con 782 kilómetros de red de distribución, 619 kilómetros de red de drenaje y 9 diques. *Distrito de Riego No. 75 Río Fuerte*, dispone de 4 plantas de bombeo, 74 pozos profundos, 2,322 kilómetros de red de distribución, 2,722 kilómetros. de red de drenaje, 2 presas derivadores y 12 diques. Finalmente el *Distrito de Riego No. 63 Guasave*, cuenta con 20 plantas de bombeo, 128 pozos profundos, 1,217 kilómetros de red de distribución, 911 kilómetros de red de drenaje, 2 Presas derivadores y 4 diques. Además para apoyar su funcionamiento los distritos de riego cuentan con una red de caminos, estructuras, casetas y edificios

La cuenca presenta dos rasgos fisiográficos principales; una zona montañosa y otra de planicie. Los rasgos montañosos se inician hacia el oriente, en dirección hacia la sierra madre occidental, que se caracteriza por presentar relieves más accidentados, donde generalmente los valles son estrechos y las corrientes poseen gradientes considerables, por lo que se considera que se encuentran en una etapa juvenil.

El clima de la cuenca hidrológica posee una gran variabilidad. En la cuenca alta prevalece un clima templado, húmedo con régimen de lluvias uniformemente repartidas, con verano fresco y largo. Mientras que en la cuenca media, el clima es semi-seco, cálido, con régimen de lluvias en verano. Mientras que la cuenca baja posee un clima seco, cálido con lluvias en verano. En los meses de julio a octubre se registran los valores de precipitación máximos que representan el 80 por ciento de la precipitación media anual.

La precipitación media anual en la Cuenca del Río Fuerte es 693 milímetros. En cuanto a la distribución temporal de la precipitación, se tienen definidos dos periodos de lluvias en la zona; las lluvias de verano y las de invierno. Las primeras son producidas por la temporada normal de lluvias y eventos meteorológicos como ciclones, los cuales se presentan con regularidad, generalmente en los meses de julio a septiembre. La segunda etapa lluviosa es producto, de los frentes fríos, durante los meses de diciembre a febrero. Por otro lado, el periodo de estiaje ocurre de marzo a mayo.

La temperatura media anual en la Cuenca del Río Fuerte es 24.15° C. La evaporación media anual en la cuenca del Río Fuerte es 2,178 milímetros, es decir mayor que la precipitación y presenta un gran desafío para la conservación del recurso hídrico. Los valores de evaporación cambian a medida que se asciende de la costa a la sierra.

Los principales ecosistemas presentan características diferentes en las partes de la cuenca. Por ejemplo en la cuenca alta dominan los suelos Kastañozems Háplicos y Lúvicos con texturas medias en pendientes quebradas. El uso del suelo es eminentemente forestal y ganadero. En la tenencia de la tierra predomina la propiedad ejidal. Mientras que la cuenca media está cubierta de selva baja caducifolia, aunque existen también algunas zonas de pastizal cultivado y pastizal inducido, así como zonas boscosas y encino en el oriente del territorio. Entre la fauna, destacan las siguientes especies conejo, liebre, coyote, zorra, tejón, armadillo, chachalaca, pato, paloma, guajolote silvestre, gato montés y venado. Finalmente la cuenca baja presenta una combinación de diferentes especies de vegetación con variedades de pastizal combinados con matorrales, destacando las áreas dedicadas a la agricultura de riego. La fauna silvestre es variada encontrándose: sapo, ninfa, sapo toro, tortuga del desierto, camaleón, lince, coyote, jabalí, liebre, conejo, tlacuache, ardilla, tortolita cola corta y paloma morada.

En la parte baja de la cuenca del Río Fuerte Sinaloa la presión para cubrir las demandas de agua ha aumentado, sobre todo en el sector agrícola, debido a que no se ha alcanzado el nivel óptimo respecto de la eficiencia en el uso y manejo del agua, tanto en los sistemas de distribución como a nivel parcelario, lo que ocasiona baja productividad agrícola, limitaciones al desarrollo socioeconómico del área, impactos negativos en los sistemas ecológicos y los consecuentes conflictos por el agua entre los usuarios. Por lo cual, el pronóstico estacional mejoraría la administración del recurso hídrico.

3.2 Generación de modelos estadísticos

En esta sección se describirán los principales procesos de regresión lineal desde lo más simple hasta llegar a modelos más complejos como son los modelos aditivos generalizados en los que se basan las técnicas estadísticas de pronóstico estacional de este proyecto. (**Fig. 3.4**)

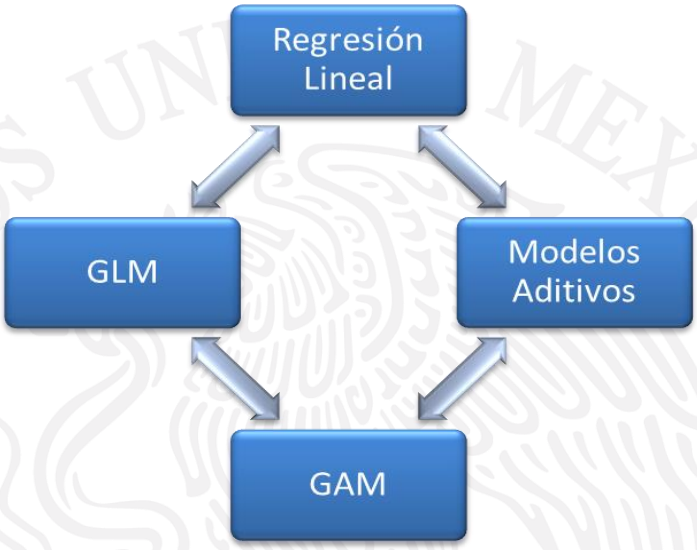


Fig. 3.4 Relaciones entre modelos de tipo de regresión.
GLM – Modelos Lineales Generalizados, GAM – Modelos Aditivos Generalizados.

3.2.1 Regresión lineal y regresión no-lineal

Los modelos estadísticos, típicamente los que se conocen como regresiones, son formalizaciones mediante las cuáles se interrelacionan entre sí varias variables aleatorias, cada una de las cuales tiene su correspondiente distribución de probabilidades. Este tema es muy general, y aquí sólo se abordará el de los modelos que se pueden formalizar como una variable, conocida como respuesta (response en inglés), definida en términos de un conjunto de variables conocidas como predictoras. Además, dado el alcance del presente texto, estos modelos se revisarán aquí solamente hasta el punto de su formulación, a partir de un conjunto de datos, conocidos como observaciones, y de su explotación, o sea su uso, sin proceder más allá al asunto de la calificación y análisis estadístico de los mismos, para lo cual hay libros especializados en el tema (p. ej. [Wood2006] y [Chambers91]).

El material de esta sección está fundamentalmente basado en el libro escrito por Santana y Mateos, 2014.

3.2.2 Modelo de regresión lineal

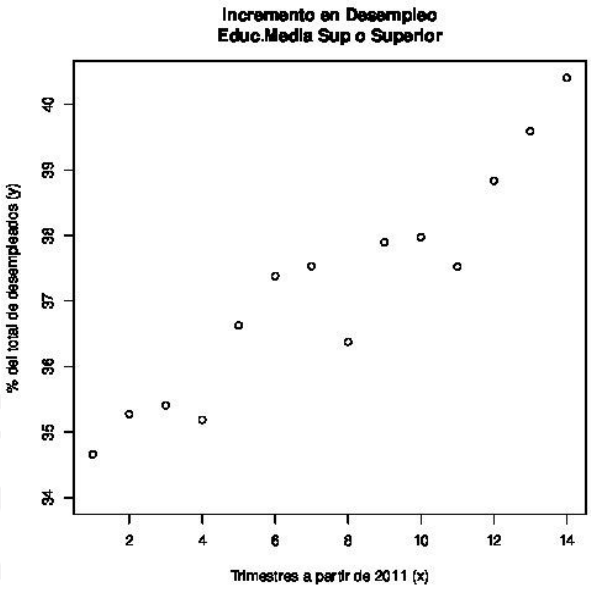
Para abordar este tema, se recurrirá a los datos sobre desempleo que presenta el sitio del INEGI en Internet: www.inegi.org.mx. Esos datos, se insertan a continuación:

```
# Porcentajes consecutivos por trimestre a partir del primer
# trimestre del año 2011, de desempleados con educación media
# superior o superior del total de desempleados en México:
prcnt <-c(
    34.6594306764, 35.2754003647, 35.40613204, 35.1855137062,
    36.6282823891, 37.3816513577, 37.5314871844, 36.3784124999,
    37.8949982178, 37.9752539821, 37.5238097329, 38.8349502588,
    39.5894958061, 40.4058337918
)
```

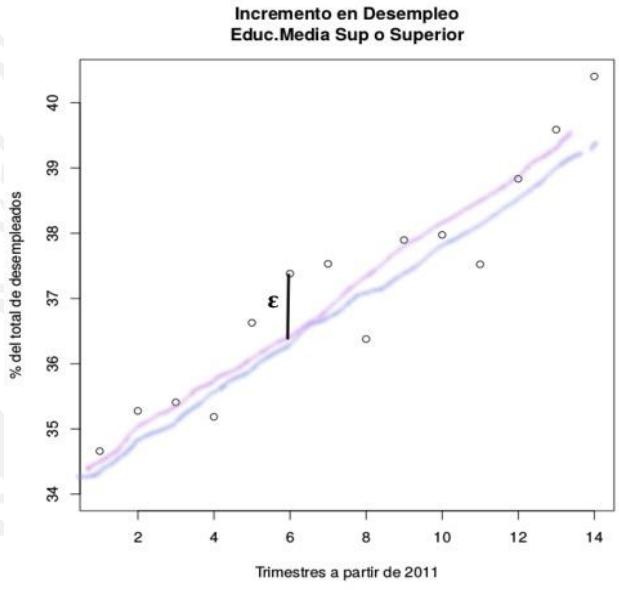
Para visualizar estos datos, se hará un gráfico de puntos, en el rango de valores que nos interesa, con el siguiente código:

```
plot(prcnt,
    ylim=c(34, max(prcnt)), # <- Rango de valores de interés
    # El título:
    main="Incremento en Desempleo\nEduc.Media Sup o Superior",
    xlab="Trimestres a partir de 2011 (x)",
    ylab="% del total de desempleados (y)"
)
```

La apariencia visual que arroja el código anterior se puede ver en la **Fig. 3.5(a)**. De ahí se intuye que uno pudiera trazar algún tipo de curva, típicamente una recta, que siguiera más o menos el camino trazado o sugerido por los puntos, como lo muestran los dos intentos de la **Fig. 3.5(b)**.



(a)



(b)

Fig. 3.5 Gráficos de desempleo en educación media superior o superior en México

Hasta aquí, aunque nuestro procedimiento, a ojo, para encontrar alguna línea que se ajuste a los datos, mismo que se ha mostrado en la **Fig. 3.5(b)**, nos da alguna idea más o menos regular de por dónde

andaría la línea en cuestión, no es de ninguna manera preciso. Supóngase, sin embargo, que la tal línea ideal existiera; entonces, para cada punto de los datos, se podría calcular la diferencia (residuo), entre el valor obtenido, en las ordenadas, en dicha línea con la misma abscisa y el del punto. Este valor se ha etiquetado en la figura para uno de los puntos con el símbolo ε. De hecho, el procedimiento para el cálculo de dichos residuos, se puede hacer para cualquier línea que se ocurra contra el conjunto de datos que se tiene. Si se suman los valores cuadrados de dichos residuos, se obtiene un valor asociado a cualquier línea candidata a ajustarse a los datos. La técnica para encontrar de entre todas las líneas candidatas posibles, la mejor u óptima, consiste en encontrar aquella cuya suma de residuos cuadrados sea la menor. Esta técnica se conoce con el nombre de regresión lineal por mínimos cuadrados y se puede estudiar a detalle en [Walpole2012] pp. 394-397 y en [Wood2006] pp. 3-12. Aquí sólo se utilizarán los resultados de la aplicación de dicho método, que se encuentran ya consignados en el lenguaje R.

En general la ecuación de cualquier recta en el espacio del problema mostrado en la **Fig. 3.5(a)**, estaría dada por:

$$y = \beta_0 + \beta_1 x \tag{3.1}$$

donde, $\beta_0$es conocida como el *intersecto* o intersección con el eje Y, y β_1 es la pendiente o inclinación de la recta. Entonces, el asunto consiste en encontrar los valores de esos coeficientes, β_0 y β_1, para la línea recta óptima, señalada en el párrafo anterior. En R, esta operación se realiza por medio de la función lm(), cuyo uso se describe a continuación con el ejemplo que se ha propuesto al principio de esta sección.

El primer paso consiste en arreglar los datos de una manera adecuada para que la función lm() los pueda manejar; esto es, se creará un data frame, con una columna correspondiente a las abscisas y otra, a las ordenadas, y sólo para usar los nombres de variables que se han empleado en la ecuación 1, se nombrarán de igual modo las columnas:

```
datos.desempleo <- data.frame(x=1:length(prcnt), y=prcnt)
head(datos.desempleo) # los primeros 6

##   x     y
## 1 1 34.66
## 2 2 35.28
## 3 3 35.41
## 4 4 35.19
## 5 5 36.63
## 6 6 37.38
```

Lo siguiente consiste en establecer la regla de dependencia entre los datos; esto es, hay que definir cuál es la variable de respuesta y cuál o cuáles son las variables *predictoras* o de estímulo. De la ecuación 3.1, se puede ver que la variable de respuesta es *y*, mientras que la única variable predictora es *x*. R tiene una manera muy particular de establecer esta relación por medio de la forma sintáctica conocida como una fórmula, y que para este caso particular se puede establecer como sigue:

```
# Para decir que 'y' depende de 'x' se
# crea una formula:
mi.formula <- y ~ x
```

Finalmente, para encontrar el ajuste lineal, se llama a la función lm(), indicando la fórmula y la fuente de los datos que se utilizarán, del siguiente modo:

```
mi.modelo <- lm(mi.formula, data=datos.desempleo)
# Ahora veamos el contenido del modelo:
mi.modelo

##
## Call:
## lm(formula = mi.formula, data = datos.desempleo)
##
## Coefficients:
## (Intercept)            x
##      34.281        0.388
```

El modelo encontrado contiene mucha información interesante desde el punto de vista estadístico y del proceso que se ha llevado a cabo para realizar el ajuste; pero, de momento, nuestro interés está en los coeficientes encontrados y que se han mostrado en la última parte del código anterior, y en la manera que se puede emplear el modelo para predecir algunos valores de la variable de respuesta *y*. Los coeficientes encontrados, en acuerdo con la ecuación 1, son $\beta_0 = 34.2813$ y $\beta_1 = 0.3879$, y con éstos, la ecuación de la línea recta encontrada sería:

$$y = 34.2813 + 0.3879x \tag{3.2}$$

En R, esta ecuación se puede implementar por medio de una función, extrayendo directamente los valores de los coeficientes del modelo como sigue:

```
mi.y <- function(x) {
    y <- mi.modelo$coefficients[1] + mi.modelo$coefficients[2]*x
    attributes(y) <- NULL
    return(y)
}
# Otra forma para extraer los coeficientes:
coef(mi.modelo)

## (Intercept)           x
##     34.2813      0.3879

# Para saber el valor que "predice" esta ecuación para una
# x = 3, correspondiente al 3er trimestre de 2011, se puede
# hacer con:
mi.y(3)

## [1] 35.45

# También se le puede pasar un vector con valores de x, así:
mi.y(9:11)

## [1] 37.77 38.16 38.55
```

Lo anterior, sin embargo, no es necesario, ya que el modelo se puede usar directamente para producir las predicciones por medio de la función predict(), como se muestra a continuación.

```
predict(mi.modelo, newdata=data.frame(x=3), type="response")

##     1
## 35.45

predict(mi.modelo, newdata=data.frame(x=9:11), type="response")
```

```
##     1     2     3
## 37.77 38.16 38.55

# Si se quisiera predecir cuál sería el porcentaje de desempleados
# con educación media-superior o superior para el 4o. trimestre
# del año 2015, ello correspondería al trimestre 20 a partir del
# primero de 2011, y se haría el cálculo con:
predict(mi.modelo, newdata=data.frame(x=20), type="response")

##     1
## 42.04
```

El modelo que se ha creado, como se ha dicho anteriormente, contiene además información estadística que se puede emplear para evaluar su *bondad*. Aunque aquí, no se entrará a todo el detalle de ella, sí se señalará cómo se puede obtener alguna de esa información, a saber, por medio de la función summary():

```
summary(mi.modelo)

##
## Call:
## lm(formula = mi.formula, data = datos.desempleo)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.0247 -0.1644  0.0563  0.3718  0.7728
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  34.2813     0.3336   102.8   <2e-16 ***
## x             0.3879     0.0392     9.9    4e-07 ***
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.591 on 12 degrees of freedom
## Multiple R-squared:  0.891,Adjusted R-squared:  0.882
## F-statistic:   98 on 1 and 12 DF,  p-value: 3.99e-07
```

Finalmente, para hacer al menos una *evaluación visual* del modelo, sería interesante producir un gráfico en el que se muestren la línea que representa el modelo junto con los datos que lo originaron. Hay varias formas en las que esto se puede hacer y en seguida se muestran dos.

```
# Primeramente se grafican los datos originales, pero, se
# extenderá el rango de las X hasta el trimestre 20, para
# observar la predicción del último trimestre del 2015:
plot(prcnt,
    ylim=c(34, 42.5), # <- Rango de valores de interés
    xlim=c(1, 20),          # <- Las x, ahora de 1 a 20
    # El título:
    main="Incremento en Desempleo\nEduc.Media Sup o Superior",
    xlab="Trimestres a partir de 2011 (x)",
    ylab="% del total de desempleados (y)",
    # Características para el desplegado:
    pch=16, col="red"
)
```

PRIMER MÉTODO:

```
# PRIMER MÉTODO: la función abline con los
# coeficientes encontrados
abline(mi.modelo$coefficients[1], # Intercepto
       mi.modelo$coefficients[2], # coef. de x
       col="blue")                 # color de línea
```

SEGUNDO MÉTODO:

```
# SEGUNDO MÉTODO: Pasando directamente el
# modelo a la función abline:
abline(mi.modelo, # El modelo
       col="blue") # color de línea
```

Cualquiera de estos dos métodos produce el gráfico que se muestra en la **Fig. 3.6**. En ese gráfico se puede corroborar también el valor predicho para el 20-*avo* trimestre a partir de 2011, es decir, el cuarto trimestre del año 2015, esto es, el 42.04 %, calculado con anterioridad.

En general, se puede establecer que una variable de respuesta, dependa de más de una variable predictora, lo cual podría dar lugar a una ecuación como la siguiente:

$$y = \beta_0 + \beta_1 x_1 + \ldots \beta_n x_n \tag{3.3}$$

donde *y* sería la variable de respuesta y, $x_1, \ldots x_n$ serían las variables predictoras. En este caso, si se suponen, por ejemplo, cuatro variables, la fórmula para establecer el modelo sería algo semejante a lo que se muestra a continuación.

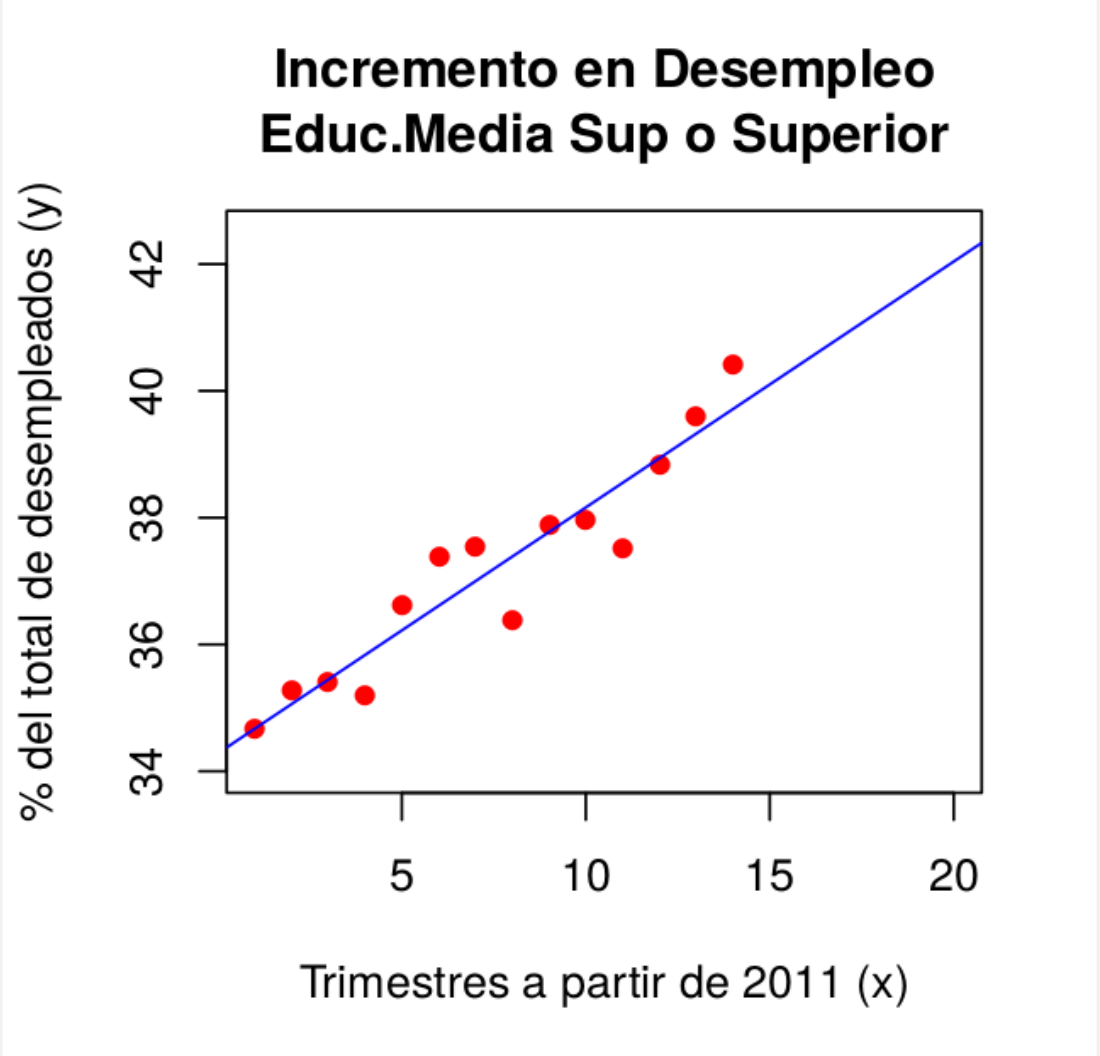


Fig. 3.6 El modelo lineal encontrado junto con los datos que lo originaron.

```
otra.formula <- y ~ x.1 + x.2 + x.3 + x.4
# las variables predictoras son:
# x.1, x.2, x.3, x.4
```

En la ecuación 3.3 no se establece exactamente la naturaleza de las variables $x_1,\ldots x_n$, de tal manera que ellas podrían ser las potencias de una misma variable y de esta manera se podría construir una expresión polinomial como la siguiente:

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \ldots \beta_n x^n \tag{3.4}$$

que puede seguirse manejando mediante un modelo lineal, y en este mismo tenor es el ejemplo que se propone a continuación.

Se ha lanzado un proyectil de un cierto edificio, y con algún instrumento de medición se han tomado medidas que describen su trayectoria. Esas medidas, tomadas de un archivo con su registro, son las que en seguida se muestran.

```
# Los datos deberán estar en el archivo:
datos.tiro <- read.table("TiroLibre-medidas.txt")
# ... esos datos son:
datos.tiro

##          x        y
## 1    0.000 15.00000
## 2    1.346 16.08960
## 3    1.686 15.80621
## 4    3.018 16.32341
## 5    4.428 15.24018
## 6    4.849 14.03364
## 7    5.882 13.74336
## 8    6.841 11.68968
## 9    7.929  9.65126
## 10   9.035  6.44392
## 11 10.307  2.25051
## 12 10.800 -0.04693
```

Con esta información, se quiere tener un modelo de la descripción de su trayectoria, y de allí determinar, más o menos, a qué altura estaría el proyectil a una distancia horizontal de 5.5 unidades.

Para tener algún tipo de apreciación de los datos que se han provisto, conviene hacer un gráfico, de la siguiente manera:

```
# Se pueden graficar especificando explícitamente
# las X y las Y, así:
#      plot(datos.tiro$x, datos.tiro$y, pch=16, col="red")
# O de manera más compacta, simplemente dando el
# data frame, fuente de los datos:
plot(datos.tiro, pch=16, col="red")
```

El resultado del código anterior se puede ver en la **Fig. 3.7**. Se observa ahí que la trayectoria no corresponde a una línea recta; de hecho, se sabe de la física que se trata de una parábola, esto es, descrita por un polinomio de segundo grado, con una ecuación semejante a la 4, de modo que hacia allá se dirigirá la solución del problema.

Los datos ya están en un *data frame* que la función lm() puede entender, de modo que nos concentraremos en la definición de la fórmula para especificar la dependencia entre la variable de respuesta y las predictoras. Para establecer la fórmula, sin embargo, se debe saber que en la sintaxis de fórmulas de R, los operadores, +, -, *, ^, :, tienen un significado distinto que el usual.

Baste por el momento, saber que el operador '+' se maneja como un separador de términos, cada uno de los cuales se compone expresiones formadas con las variables predictoras; y el operador '-', se usa para eliminar términos. Si lo que se quiere es introducir expresiones en las que los operadores anteriores tengan su significado usual, ellas deben estar encerradas en el constructor sintáctico I( ). Como ese es el caso de lo que se desea expresar, nuestra fórmula de dependencia se construye de la siguiente manera:

```
formula.tiro <- y ~ x + I(x^2)
# El significado de x^2 dentro de I( )
# se toma 'textualmente', esto es, su
# significado es "equis cuadrada".
```

Así, el modelo *lineal*, de acuerdo con esa fórmula, ajustado a los datos provistos se construye como sigue:

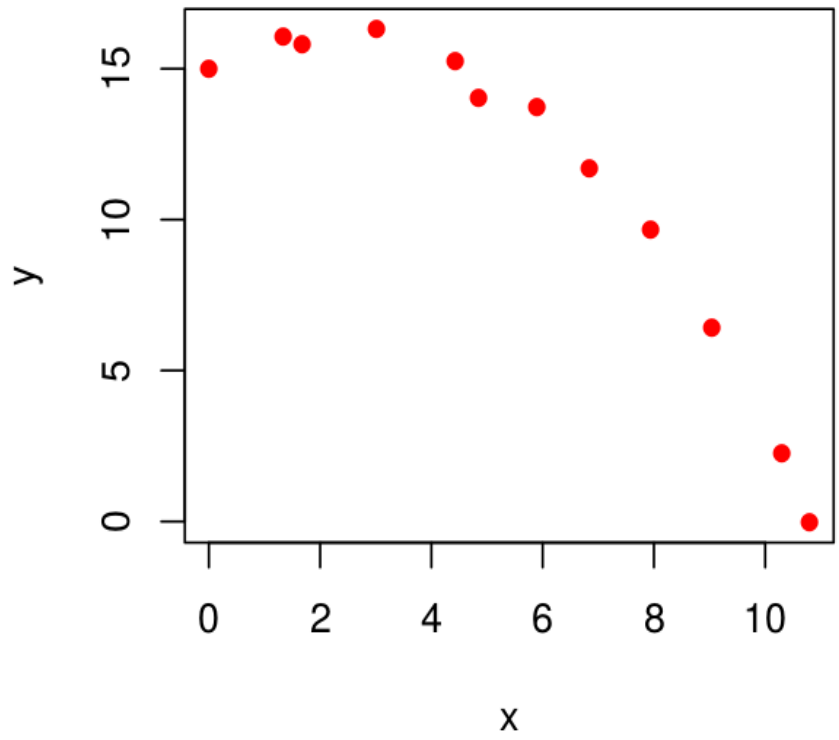


Fig. 3.7 Medidas del lanzamiento de un proyectil.

```
modelo.tiro <- lm(formula.tiro, data=datos.tiro)
# Veamos la información básica del modelo:
modelo.tiro

##
## Call:
## lm(formula = formula.tiro, data = datos.tiro)
##
## Coefficients:
## (Intercept)            x       I(x^2)
##      14.902        1.069       -0.224
```

De esta manera, la ecuación de la trayectoria del proyectil ajustada a los datos estaría dada por:

$$y = 14.902 + 1.069x - 0.224x^2 \tag{3.5}$$

Tanto para el uso del modelo, como para graficar la curva junto con los datos que lo originaron, es conveniente hacer una función basada en la función predict() del lenguaje, de la manera siguiente:

```
func.tiro <- function(x) {
    # El argumento x puede ser un solo número o un vector de
    # números
    predict(modelo.tiro, newdata=data.frame(x=x), type="response")
}
```

Y el problema de saber la altura del proyectil a 5.5 unidades, junto con otros valores, se resuelve así:

```
# Altura a 5.5 unidades horizontales:
func.tiro(5.5)

##  1
## 14

# Si se quiere conocer las alturas predichas por el modelo
# para las distancias 3.5, 4.3 y 8.2:
func.tiro(c(3.5, 4.3, 8.2))

##      1      2      3
## 15.900 15.357  8.603
```

Finalmente, si se quiere graficar la curva correspondiente al modelo, junto con los datos, se hace agregando al gráfico de la **Fig. 3.7**, la salida del código que se presenta a continuación. En este código se está añadiendo, además, sólo para comparar, lo que sería el resultado si los datos se hubieran ajustado a una línea recta.

```
# La curva ajustada:
curve(func.tiro, lwd=2, col="blue", add=T)
# Hagamos el modelo de una recta:
mod.recta <- lm( y ~ x, data=datos.tiro)
# Dibujémosla en el gráfico también:
abline(mod.recta) # va en negro, el default
# ---------------
# Pondremos una leyenda apropiada para identificar
# los objetos en el gráfico:
legend(
   "bottomleft", # Donde va la leyenda
   legend=c("modelo parabola", "modelo recta"), # textos
   lwd=c(2,1), # anchos de línea
   col=c("blue", "black") # colores para cada modelo
)
```

El resultado de este código, añadido a la figura anterior, se puede apreciar en la **Fig. 3.8**.

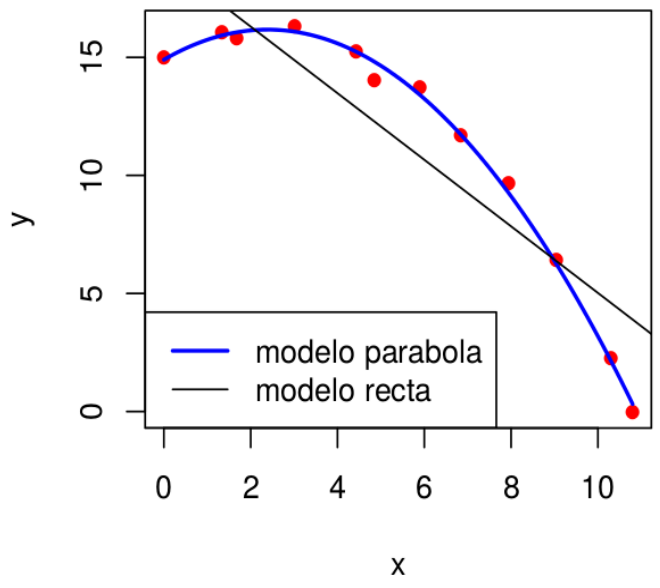


Fig. 3.8 Dos modelos estadísticos ajustados con un mismo conjunto de datos.

3.2.3 Modelos lineales generalizados

En los modelos que se han revisado en la sección anterior, se parte de la suposición de que la distribución de los residuos tiene una distribución normal, por lo menos de manera aproximada. Los residuos son la diferencia entre el valor observado y el valor obtenido por el modelo. Por otra parte, la variable aleatoria que se ha observado es continua y puede tomar cualesquiera valores. En muchos problemas, sin embargo, la variable de respuesta que se desea modelar no obedece a dichas características. Pensemos, por ejemplo, en variables que pudieran tener una repuesta binaria: "infectado" o "no infectado", "aprobado" o "reprobado", etc. Por supuesto que este tipo de variables no tienen una distribución normal.

Antes de entrar en la materia de los modelos lineales generalizados, daremos un poco de formalidad a los modelos lineales revisados en la sección anterior.

En las figuras 2 y 4 se muestran tanto los datos *experimentales*, en rojo, como las *curvas* resultantes del ajuste, en azul. En ambos casos, la curva o línea azul representa la media, μ, de la variable aleatoria, Y, y, por consiguiente, cada uno de los valores de la variable se podría expresar como sigue:

$$y_i = \mu + \varepsilon_i = \beta_0 + \beta_1 x_1 + \ldots \beta_n x_n + \varepsilon_i \tag{3.6}$$

donde, en la terminología de los modelos estadísticos, los ε_i son conocidos como los residuos del modelo[1]. Si se eliminan éstos de la ecuación anterior, se tiene que la media es el objeto o variable que está definida por la expresión lineal, como se muestra:

$$\mu = \beta_0 + \beta_1 x_1 + \ldots \beta_n x_n \tag{3.7}$$

Los modelos lineales generalizados parten de una expresión semejante a la anterior. Pensemos por un momento que dicha expresión no es la media, sino una variable cualquiera, η, que depende linealmente de las variables predictoras, así:

$$\eta = \beta_0 + \beta_1 x_1 + \ldots \beta_n x_n \tag{3.8}$$

Si ahora se piensa que de alguna manera la variable aleatoria, Y, y por consiguiente su media, μ, dependen de η, que se conoce como el *predictor lineal*, eso se puede expresar de la siguiente manera:

$$\mu = m(\eta), \eta = m^{-1}(\mu) = L(\mu) \tag{3.9}$$

1 Son precisamente estos residuos los que se presupone, como se dijo anteriormente, obedecen a una distribución normal de probabilidades.

Aquí, L se conoce como la función de liga, o *link*, ya que es la que establece una liga entre el predictor lineal, η, y la media real de la variable aleatoria, *Y*. Por su parte, la variable aleatoria, *Y*, se supone con una distribución perteneciente a la familia de las distribuciones exponenciales, entre las que se encuentran: la Normal, la Gamma, la Binomial, la Poisson, y algunas otras más.

ID	Horas-curso	Grupo	Aprobados	Proporción
A	15.5	20	2	0.1
B	21.8	20	4	0.2
C	28.1	21	6	0.286
D	34.5	19	9	0.474
E	40	25	15	0.6
F	40.9	21	18	0.857
G	47.3	20	18	0.9
H	53.6	21	21	1
I	60	20	19	0.95

Tabla 3.1 Caso de los estudiantes en programa piloto.

Para revisar este tipo de modelos estadísticos, se propone aquí un ejemplo en el terreno de lo que se conoce como *regresión logística* (*logit*).

- Ejemplo de regresión logística

Considérese el siguiente caso:

En un sistema escolar se han diseñado cursos remediales de diversa extensión en horas de curso. Los cursos se aplicaron en un programa piloto a nueve grupos de estudiantes previamente reprobados durante los cursos normales del sistema. Los resultados del programa piloto se muestran en la **Tabla 3.1**.

En este ejemplo, lo primero que resalta es que el resultado importante se expresa en términos binarios, esto es, como *aprobado* o *no-aprobado*, que, en la terminología de las distribuciones de probabilidades que manejan este tipo de resultados, se traduce como *éxito* (1) o *fracaso* (0) de la prueba. Aquí pues, lo interesante es modelar la *probabilidad* de éxito o fracaso en función de un conjunto de variables predictoras, que para el ejemplo se podría reducir a la única variable *horas-invertidas-en-el-curso*, representada por la columna *Horas-curso* en la **Tabla 3.1**. La probabilidad, sin embargo, tiene un valor que fluctúa entre 0 y 1, mientras que los modelos estadísticos lineales, revisados en la primera sección de este capítulo, están diseñados de tal manera que las variables de respuesta pueden tomar valores en todo el rango de los números reales. Es aquí, justamente donde la utilidad de la definición de una función de liga, introducida en la fórmula de la ecuación 9, se manifiesta, ya que ella establecerá una transformación entre un espacio y el otro.

Para abordar este tema, se recurre primeramente a la noción de la esperanza matemática condicional. En general, la esperanza matemática de una variable aleatoria se traduce en la media de la variable; esto es, E(Y) = μ. Por otra parte, la esperanza condicional, denotada como E($Y \mid X$) , es la esperanza de Y dado que ocurrió X, donde, para los modelos lineales vistos anteriormente, X podría representar el conjunto de las variables predictoras, en cuyo caso se tendría que $E(Y/X) = \beta_0 + \beta_1 x_1 + \ldots + \beta_n x_n$.

Ahora bien, si Y es una variable aleatoria binaria que solamente toma los valores 1 o 0, la expresión para la esperanza matemática condicional estaría dada por:

$$E(Y/X) = Pr(Y = 1/X) \tag{3.10}$$

Así, la meta es modelar la probabilidad de éxito (Y=1), dado X, el conjunto de variables predictoras. En el caso particular del ejemplo propuesto, sólo hay una variable predictora, a saber, las *Horas-curso*; esto es, la variable x, denotará el número de horas invertidas en el curso. De la **Tabla 3.1**, se puede observar que la proporción de éxito (número de estudiantes aprobados), en general, tiende a aumentar a medida que se incrementa el número de horas en invertidas en el curso.

Para analizar el problema, consideremos cualquier renglón de la tabla de datos, digamos el tercero. Aquí, al asistir a 28.1 horas de curso, de 21 estudiantes que conformaban el grupo, 6 resultaron aprobados. La modelación de este tipo de resultados, se hace mediante la distribución de probabilidades binomial. En este tipo de distribución, un experimento consiste de n pruebas o ensayos independientes, cada uno de los cuales tiene dos resultados posibles: *éxito* (1) o *fracaso* (0), y la probabilidad de éxito, p, en cada prueba permanece constante. Para el renglón que estamos considerando, n = 21, y la frecuencia de éxitos se puede tomar como la probabilidad, así, p = 6/21 = 0.286. En estas condiciones la función de probabilidad binomial, en general y en particular para este caso, está dada por:

$$Pr(y = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

$$Pr(y = k) = \binom{21}{k} 0.286^k (1-0.286)^{21-k} \tag{3.11}$$

donde, $\binom{n}{k} = \frac{n!}{k!(n-k)!}$, se refiere a las combinaciones de n en k; esto es, n elementos tomados de k en k, y donde además k = 0, 1, . . . 21. En R, la función binomial de probabilidades se calcula con la función dbinom(). Como un ejemplo ilustrativo, incluimos aquí el código que permite hacer un gráfico de las probabilidades para los datos del tercer renglón de **Tabla 3.1**, y cuyo resultado se muestra en la **Fig. 3.9**.

```
# Creador de funciones de probabilidades binomiales
creaFBinom <- function(n,p) function(k) dbinom(k, n, p)
# Para el caso del ejemplo:
n <- 21
p <- 6/21 # frecuencia aprox= probabilidad
ffb <- creaFBinom(n, p)
# ffb(k) sería la función binomial para una k dada
# .. para el caso de todas las k de nuestro interés:
k <- 0:n # Esto es: 0,1,..,21
# Para este caso las probablididades son:
ffb(k)

##  [1] 8.537e-04 7.171e-03 2.868e-02 7.267e-02 1.308e-01
##  [6] 1.779e-01 1.898e-01 1.626e-01 1.139e-01 6.578e-02
## [11] 3.157e-02 1.263e-02 4.210e-03 1.166e-03 2.665e-04
## [16] 4.974e-05 7.461e-06 8.778e-07 7.803e-08 4.928e-09
## [21] 1.971e-10 3.755e-12

# Y la media (esperanza matemática) está dada por
# la suma de los productos de cada valor por su
# respectiva probabilidad, esto es:
(media <- sum(k*ffb(k)))

## [1] 6

# .. y la probabilidad asociada a este valor es:
ffb(media)

## [1] 0.1898
# Ahora procedemos a graficar esto con:
barplot(ffb(k), names.arg=k, xlab="k", ylab="Pr(k)")
```

El significado tanto de la fórmula, como del gráfico presentado en la **Fig. 3.9** es que la altura de cada barra representa la probabilidad de que *k* estudiantes (el valor de *k* correspondiente a la barra), de los 21 que conforman el grupo, salieran aprobados en el programa piloto. No es una sorpresa que justamente de acuerdo con la función, $k = 6$, representa la media de la variable, y tiene una probabilidad de 0.1898.

Ahora bien, cada renglón de la **Tabla 3.1**, tiene su propia función de probabilidades binomial y su propia representación gráfica como la mostrada en la **Fig. 3.9**. Esto es, básicamente, en la ecuación 3.11, el parámetro *p* (probabilidad de éxito) dependería del *renglón* de la tabla que se quisiera modelar, y éste, a su vez, identifica el caso del número de horas invertidas en el curso, que es, para el ejemplo, la única variable predictora, llamémosle x, que se usará. En otras palabras $p = p\,(\,x\,)$, con lo que la expresión citada sería semejante a:

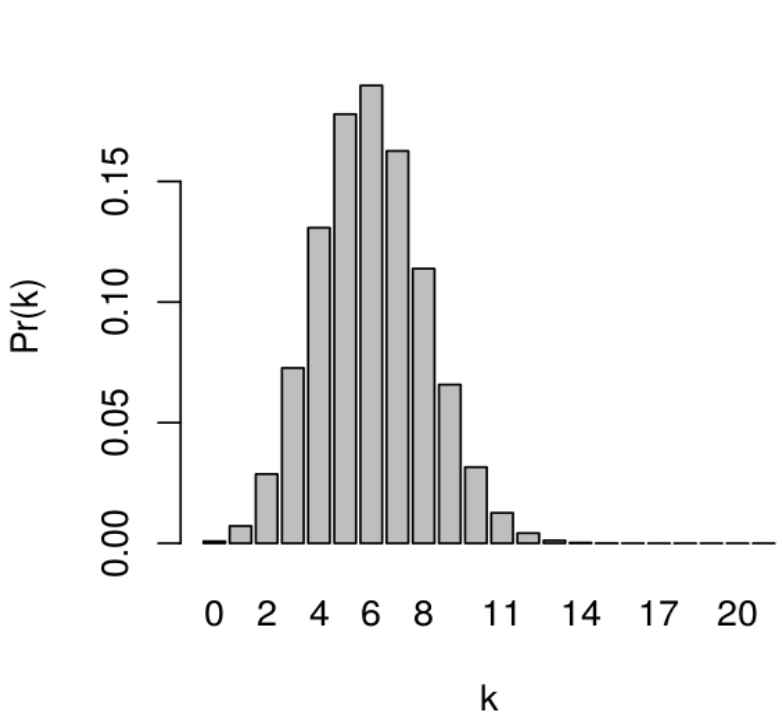


Fig. 3.9 Función de probabilidades binomial: $n = 21$, $p = 6/21$

$$Pr(y = k) = \binom{n}{k} p(x)^k (1 - p(x))^{n-k} \tag{3.12}$$

Por otra parte, todos los valores de estas probabilidades, como se ha dicho antes, fluctúan entre 0 y 1. Se requiere por tanto de funciones que mapeen del dominio de todos los números reales, a los valores comprendidos en el intervalo entre 0 y 1, y viceversa. Unas funciones que sirven a este propósito y que son ampliamente usadas, son la función *logística* y su inversa, la función *logit*, cuyas definiciones respectivas se muestran a continuación, y cuyos gráficos se muestran en la **Fig. 3.10**.

$$logistic(t) = \frac{e^t}{1 + e^t} = \frac{1}{1 + e^{-t}} \tag{3.13}$$

$$logit(r) = log\left(\frac{r}{1-r}\right) = log(r) - log(1-r) = -log\left(\frac{1}{r} - 1\right) \tag{3.14}$$

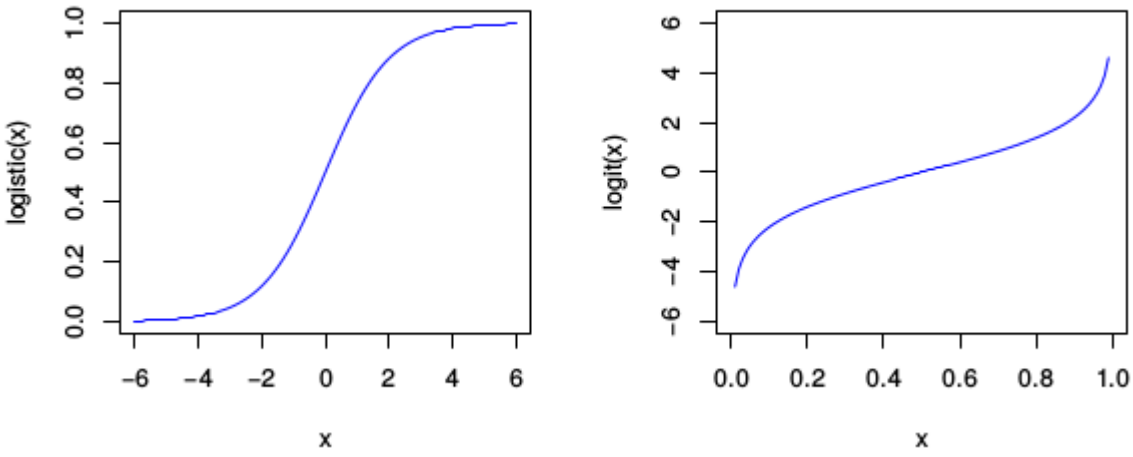


Fig 3.10. Las funciones *logística* y su inversa *logit*

¿Cómo intervienen estas expresiones en el tema? Primeramente observemos en la ecuación 3.12 que la probabilidad *p* depende de *x*. Pero, mientras que la variable *x* toma valores arbitrarios en el espacio de los números reales, por ejemplo 15.5, 21.8, 28.1, etc., *p*(*x*) toma valores entre 0 y 1, de modo que se necesita realizar un mapeo del tipo establecido por la función logística de la ecuación 3.13. En efecto, el mapeo en cuestión es el siguiente:

$$p(x) = logistic(\beta_0 + \beta_1 x) = \frac{e^{\beta_0+\beta_1 x}}{1 + e^{\beta_0+\beta_1 x}} \tag{3.15}$$

y, manipulando algebraicamente la ecuación, se puede llegar a:

$$log\left(\frac{p(x)}{1-p(x)}\right) = \beta_0 + \beta_1 x$$

$$logit\,(p(x)) = \beta_0 + \beta_1 x \tag{3.16}$$

En esta expresión, el cociente $\frac{p(x)}{1-p(x)}$ se conoce como la *razón de oportunidades* p^2, o *momios* de y es la probabilidad de éxito dividido entre la probabilidad de fracaso de una cierta prueba. Nótese además que el lado derecho en las dos expresiones de la ecuación 3.16 corresponde al predictor lineal que se persigue, y así, *logit*(*p* (*x*)) , se constituye en la función de liga para el caso.

2 En inglés *odds ratio*.

El problema ahora consiste en encontrar β_0 y β_1 a partir de expresiones como las mostradas en las ecuaciones 3.12, 3.15 y 3.16, y se resuelve por medio de técnicas tales como el método de la *estimación de la máxima verosimilitud*[3].

Para introducir brevemente el método de estimación de la máxima verosimilitud, abundaremos en algunos detalles del problema propuesto. La **Tabla 3.1**, caracteriza a varios grupos de estudiantes, que identificaremos con las letras A, B, C, D, E, F, G, H, e I. Supongamos que se sabe que un pequeño subgrupo de 15 estudiantes, entre los que hay 11 que aprobaron el curso, harán una fiesta; sin embargo, no se sabe de cuál de los grupos originales proviene el subgrupo. El problema es entonces determinar de cuál de estos grupos hay más posibilidades que provenga el subgrupo. Para medir esto, se calculará un indicador, *L*, al que denominaremos verosimilitud, con los datos proporcionados de n=15 y k=11. La fórmula, semejante a la de la ecuación 11, solamente que haciendo explícito que *L* está en función de *p*, es la siguiente:

$$L(p) = \binom{n}{k} p^k (1-p)^{n-k} \tag{3.17}$$

En R, podemos resolver el asunto, para cada uno de los renglones de la tabla, así:

```
# Primeramente leemos la tabla:
curso <- read.csv("CursoPiloto.csv", header=T)
# Veamos un fragmento de la tabla:
head(curso)

##   ID Horas Grupo Aprobados Proporcion
## 1  A  15.5    20         2      0.100
## 2  B  21.8    20         4      0.200
## 3  C  28.1    21         6      0.286
## 4  D  34.5    19         9      0.474
## 5  E  40.0    25        15      0.600
## 6  F  40.9    21        18      0.857

# Creador de funciones de probabilidades binomiales,
# .. sólo que ahora la función resultante tiene como
# .. argumento p
creaFBinom2 <- function(n,k) function(p) dbinom(k, n, p)
# En el caso que nos ocupa:
```

3 En inglés *likehood*.

```
n <- 15;  k <- 11
# .. y la función de Verosimilitud L es:
L <- creaFBinom2(n, k)
# La columna tt$Proporcion puede ser tomada como la
# probabilidad
# Agreguemos a 'curso' una columna con L para cada caso:
curso$L <- L(curso$Proporcion)
curso

##   ID Horas Grupo Aprobados Proporcion         L
## 1  A  15.5    20         2      0.100 8.956e-09
## 2  B  21.8    20         4      0.200 1.145e-05
## 3  C  28.1    21         6      0.286 3.715e-04
## 4  D  34.5    19         9      0.474 2.836e-02
## 5  E  40.0    25        15      0.600 1.268e-01
## 6  F  40.9    21        18      0.857 1.045e-01
## 7  G  47.3    20        18      0.900 4.284e-02
## 8  H  53.6    21        21      1.000 0.000e+00
## 9  I  60.0    20        19      0.950 4.853e-03
```

Del resultado anterior, se puede ver que el grupo con el máximo valor de L es el identificado con la letra E, para el que se encontró $L(0.6)=0.1268$, y es, por tanto, el que tiene más posibilidades de ser el origen del subgrupo de la fiesta.

Ahora bien, consideremos el problema en el que las probabilidades, p, no son fijas sino que dependen de x, de acuerdo con una expresión como la mostrada en la ecuación 3.15, y considerando que los resultados de cada uno de los grupos son independientes entre ellos, se puede aplicar la regla del producto de probabilidades y así, el indicador de verosimilitud estaría dado por:

$$L = \binom{n_1}{k_1} p(x_1)^{k_1} (1 - p(x_1))^{n_1 - k_1} \times \ldots \times \binom{n_m}{k_m} p(x_m)^{k_m} (1 - p(x_m))^{n_m - k_m} \quad (3.18)$$

donde,

$$p(x_i) = \frac{e^{\beta_0 + \beta_1 x_i}}{1 + e^{\beta_0 + \beta_1 x_i}}$$

para i = 1, 2, ..., m.

No existe una forma directa de encontrar los valores de β_0 y $\beta_1$que maximizan la expresión dada en la ecuación 18, de modo que se resuelve por métodos de aproximaciones sucesivas, como el método de Newton-Raphson, o el método iterativo de mínimos cuadrados re-ponderado. Afortunadamente R, ya incorpora en la función glm(), un método de esos. El tipo de especificación que se usa con esta función es bastante similar al empleado con la función lm(), revisada al principio de este capítulo, con las siguientes distinciones para el caso de distribución binomial, que es el que nos ocupa:

- Se debe especificar distribución que se aplica mediante el argumento family de la función: family=binomial.

- En la fórmula y los datos, la variable de respuesta, para el caso de la distribución binomial, se puede especificar de tres maneras:

1. Como un *vector*. En este caso, el sistema entendería que se trata de datos binarios y por lo tanto el vector tendría sólo valores 0 y 1.
2. Como una *matriz* de dos columnas. En este caso, se entiende que la primera columna contiene el número de éxitos para la prueba y la segunda columna, el número de fracasos.
3. Como un *factor*. En este caso, el primer nivel (o categoría) del factor, se entiende como fracaso (0), y todos los otros como éxito (1).

- La familia binomial, por omisión toma como función de liga la función logit, si se quisiera especificar otra de las disponibles, por ejemplo la función probit, se haría así: family=binomial(link=probit).

Una vez establecido lo anterior, procedemos a resolver el problema en R, de la siguiente manera:

```
reprobados <- curso$Grupo-curso$Aprobados
aprobados <- curso$Aprobados
horasCurso <- curso$Horas
(exito <- aprobados/(aprobados+reprobados))

## [1] 0.1000 0.2000 0.2857 0.4737 0.6000 0.8571 0.9000 1.0000
## [9] 0.9500

# Aquí optamos por especificar la variable de respuesta
# como una matriz de dos columnas:
formula <- cbind(aprobados,reprobados) ~ horasCurso
mm <- glm(formula, family=binomial)
summary(mm)

##
## Call:
## glm(formula = formula, family = binomial)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -1.240  -0.554   0.343   0.439   1.547
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -4.7056     0.7387   -6.37  1.9e-10 ***
## horasCurso    0.1407     0.0201    7.00  2.6e-12 ***
```

```
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 95.6146  on 8  degrees of freedom
## Residual deviance:  6.9977  on 7  degrees of freedom
## AIC: 34.24
##
## Number of Fisher Scoring iterations: 4

# Hagamos una función para predecir con el modelo
ff <- function(x) predict(mm, newdata=data.frame(horasCurso=x),
                          type="response")
# De acuerdo con el modelo, las probabilidades de éxito
# para cursos de 30 y 50 horas serían:
ff(c(30,50))

##      1      2
## 0.3813 0.9113
```

El resultado visual del código anterior se puede ver en la **Fig. 3.11**.

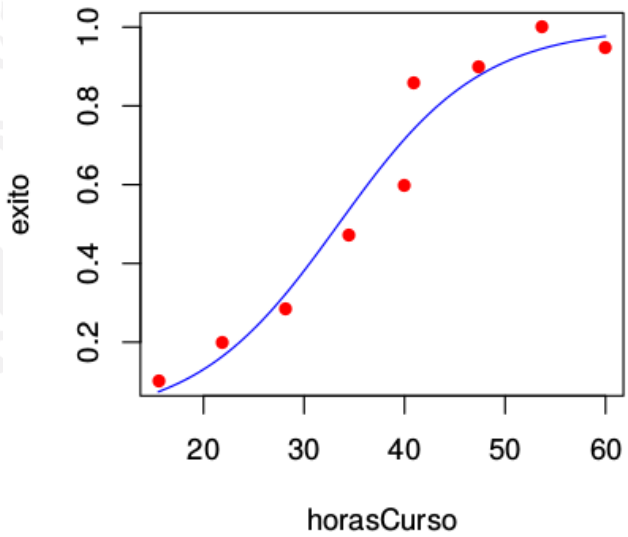


Fig. 3.11 Resultados del problema de regresión logística resuelto con la función glm().

3.2.4 Modelos aditivos generalizados

Los Modelos Aditivos Generalizados son modelos de regresión similares a los GLM pero en los que $Y_i|x_i$ sigue una distribución de la familia exponencial en la que la media viene dada por

$$\mu = l^{-1} (\Sigma \beta i \, fi \, xi))$$

Las f_i son funciones suaves que permiten reflejar efectos no lineales de las variables x_i sobre la variable Y. Una solución sencilla para tener en cuenta estos efectos no lineales hubiese sido incorporar en el predictor lineal términos cuadráticos, cúbicos, etc., de las variables explicativas (es decir, términos de la forma x^2, x^3, etc.) de forma que el predictor deje de ser lineal y se convierta en un polinomio de las variables explicativas. Sin embargo, esta solución plantea problemas bien conocidos derivados del hecho de que los polinomios son funciones de soporte no acotado (su dominio se extiende sobre todo el eje real), de manera que, en general, cualquier intento de mejorar su ajuste en un punto determinado se consigue a expensas de empeorarlo en otros puntos muy alejados. Esto supone que los polinomios pueden proporcionar una solución adecuada cuando se busca un buen ajuste en el entorno de un punto concreto, pero no sobre todo un intervalo.

Una alternativa ampliamente utilizada consiste en emplear *splines*. Los *splines* son polinomios definidos sobre intervalos y que toman valor nulo fuera de esos intervalos. Con esto se consigue graduar el ajuste de la regresión de forma que los cambios que se produzcan para mejorar ese ajuste tengan un efecto local y no se extiendan más allá de los intervalos en los que están definidos los *splines* involucrados en cada caso. La **Fig. 3.12** muestra el ejemplo de un *spline* y un conjunto de ellos:

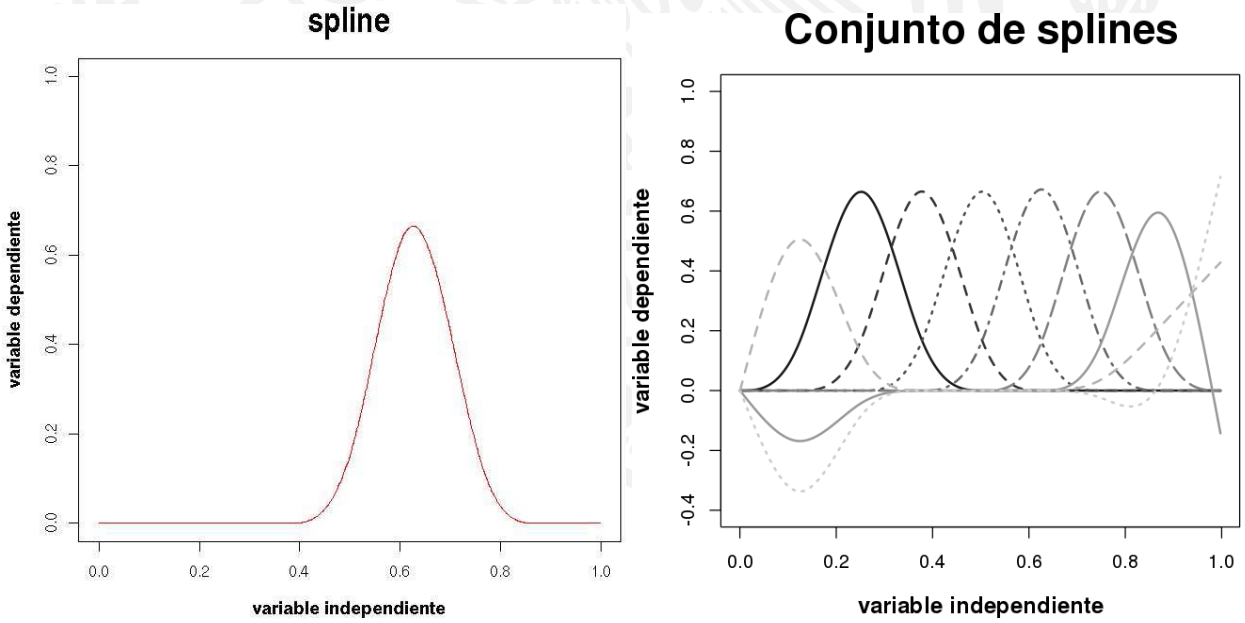


Fig. 3.12 Diagrama esquematizando un *spline* y un conjunto de ellos.

Los intervalos sobre los que se definen los *splines* vienen determinados por un conjunto de puntos denominados *knots*. Un conjunto de *q - 2 knots* determinan *q - 1* intervalos, que pueden estar contenidos dentro de un intervalo acotado *[a, b]* o extenderse hasta -∞ y +∞. Habitualmente se utilizan *splines* cúbicos, que son polinomios de grado 3, con lo que cada uno de ellos queda determinado por 4 parámetros. En consecuencia el espacio vectorial de los *splines* formados por combinaciones lineales de polinomios cúbicos definidos sobre cada uno de los *q - 1* intervalos determinados por *q - 2 knots* y nulos en el resto de intervalos, es un espacio de dimensión *4(q - 1)* . Sin embargo, se introducen una serie de restricciones que reducen la dimensión de este espacio. Estas restricciones consisten en que se exige que los *splines* y sus primeras y segundas derivadas sean continuas en los *q - 2 knots*. Son entonces *3(q - 2)* condiciones que hacen que la dimensión del espacio vectorial de los *splines* cúbicos definidos sobre el intervalo *[a, b]* sea *4(q -1) - 3(q - 2) = q + 2*. Para los *splines* denominados "naturales" también es habitual exigir que su derivada segunda se anule en los extremos del intervalo *[a, b]*, lo que da lugar a otras dos restricciones adicionales que hacen que finalmente la dimensión del espacio vectorial de los *splines* naturales cúbicos definidos por *q - 2 knots* sea *q*.

Existen muchas formas distintas de expresar una base de tal espacio vectorial. Una es, por ejemplo, la denominada base de potencias truncadas (Hastie *et al*., 2008), que se obtiene a partir de potencias de funciones "parte positiva" de la forma $(X - \xi_j)$, en donde ξ_j es un *knot*, y que se definen como 0 si su argumento es negativo y como $X - \xi_j$ si este argumento es positivo.

En cualquier caso, si denominamos *sk* a las funciones de la base del espacio de los *splines* cúbicos naturales definidos en un determinado intervalo, tendremos que cualquier función suave definida sobre ese intervalo se puede aproximar con una combinación lineal de las *sk*. De acuerdo con esto, las funciones *fi* del predictor lineal de los modelos GAM se podrán expresar como combinaciones lineales de las funciones *sk*, y la estimación del predictor lineal se reduce a la estimación de los coeficientes de las combinaciones lineales de las funciones *sk*. Esto equivaldría a minimizar la cantidad $\|2Y - X\beta\|$, donde los valores de X serán los que adopten las funciones *sk* en las observaciones disponibles.

Planteada de esta manera la estimación de un GAM sería esencialmente idéntica a la estimación de un GLM. Sin embargo, existe una complicación adicional derivada de la mayor o menor "suavidad" (o "rugosidad") exigida a las funciones que se estiman. Esta rugosidad depende fundamentalmente del número de *knots* que se utilicen, pues cuanto mayor sea este número mayor la rugosidad de las funciones estimadas, y un número de *knots* elevado conducirá a un sobreajuste.

Considerando que una medida global de la rugosidad de una curva viene dada por la expresión

$$\int [f''(x)]^2 \, dx \, ,$$

Conviene introducir en el proceso de estimación una penalización de esta rugosidad para evitar el sobreajuste. El objetivo del proceso de estimación será entonces minimizar la cantidad siguiente:

$$\|2Y - X\beta\| + \lambda \int [f''(x)]^2 \, dx$$

El parámetro λ se denomina parámetro de alisado y controla el grado de alisado (o de rugosidad) de la función estimada. Un valor bajo de λ hace que la rugosidad penalice poco, lo que conducirá a un sobreajuste (se ajusta el ruido además de la señal), mientras que un valor alto de λ hará que se penalice mucho esa rugosidad, con lo que se estimará una línea recta. El valor adecuado para este parámetro se suele seleccionar recurriendo a alguna variante de *cross-validation*, es decir, se van extrayendo observaciones de una en una, se ajusta el modelo para las observaciones restantes, se predice la respuesta correspondiente a la observación eliminada y se compara con el valor real omitido. Combinando los resultados para todas las observaciones se obtiene una medida global de ajuste del modelo. Finalmente se selecciona el valor de λ para el cual ese ajuste es mejor. En la práctica se suelen usar como medidas de ajuste las cantidades denominadas UBRE (*Un-Biased Risk Estimator*) o GCV (*Generalized Cross Validation*), según que el parámetro de dispersión sea conocido o estimado, respectivamente (Wood, 2006).

Por otra parte se puede ver con facilidad que la rugosidad global puede ser expresada como una forma cuadrática de los parámetros β:

$$\int [f''(x)]^2 \, dx = \beta^t S \beta,$$

con lo que el estimador de estos parámetros se puede expresar como:

$$\beta = (X^t X + \lambda S)^{-1} X^t Y$$

y la *hat matrix A* que permite obtener los valores estimados a partir de los observados ($\hat{\mu} = AY$) será $A = X\hat{\beta}$.

Al igual que en un modelo lineal convencional la *hat matrix* es una matriz de proyección cuya traza proporciona la dimensión del espacio sobre el que se proyecta, y por lo tanto el número de grados de libertad del modelo, se definen en los GAM los grados de libertad del modelo como *tr(A)*. El número de grados de libertad será una medida de la complejidad (y rugosidad) del modelo. Un modelo con muchos grados de libertad será un modelo muy complejo, con muchos parámetros y sobreajustado, mientras que uno con pocos grados de libertad probablemente deje sin ajustar aspectos relevantes de las observaciones. Conviene resaltar también que esos grados de libertad totales se pueden descomponer como suma de grados de libertad correspondientes a cada una de las variables del modelo, o con mayor precisión, a cada una de las funciones *fi* que aparecen en el predictor lineal.

3.2.5 Algunas métricas importantes para la evaluación del pronóstico

- Error Medio Absoluto (MAE)

En estadística, el error absoluto medio (MAE) es una cantidad que se usa para medir qué tan cerca están los pronósticos o predicciones son para los valores observados. El error absoluto medio viene dado por

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}|f_i - y_i| = \frac{1}{n}\sum_{i=1}^{n}|e_i|\,.$$

Como su nombre indica, el error medio absoluto es un promedio de los errores absolutos $e_i = |f_i - y_i|$, donde f_i es la predicción y el y_i valor observado. Tenga en cuenta que las formulaciones alternativas pueden incluir frecuencias relativas como factores de peso.

El error absoluto medio es una medida común de error de pronóstico en el análisis de series de tiempo, donde los términos "significan desviación absoluta" se utilizan a veces en confusión con la definición más estándar de la desviación media absoluta.

- La Eficiencia Nash-Sutcliffe (NSE)

Es una estadística normalizada que determina la magnitud relativa de la varianza residual ("ruido") en comparación con la variación de datos de medida ("información") (Nash y Sutcliffe, 1970).

NSE indica qué tan bien la gráfica de los datos observados frente a los se ajusta a la línea 1:1. El valor de la eficiencia Nash-Sutcliffe va desde -∞ a 1. En esencia, entre más cercano a 1, es más preciso es el modelo.

- NSE = 1, corresponde a una combinación perfecta de modelo a los datos observados.
- NSE = 0, indica que las predicciones del modelo son tan precisos como la media de los datos observados,
- -∞<NSE <0, indica que la media observada es mejor predictor que el modelo.

- Índice de Habilidad Heidke (HSS)

El índice de habilidad Heidke está en el formato de índice de habilidad de costumbre,

Habilidad = (valor del score - score para el pronóstico estándar) / (score perfecta - score para la previsión estándar)

Para el HSS, el "score" es el número correcto o la proporción correcta. El "pronóstico estándar" es por lo general el número correcto por casualidad o la proporción correcta por casualidad. Por lo tanto usando la proporción correcta,

Event forecast	Event observed		
	Yes	No	Marginal total
Yes	a	b	a + b
No	c	d	c + d
Marginal total	a + c	b + d	a + b + c + d =n

$$HSS = \{(a + d) / n - [(a + b) (a + c) + (b + d) (c + d)] / n^2\} / \{1 - [(a + b) (a + c) + (b + d) (c + d)] / n^2\}$$

Esto se puede simplificar en,

$$HSS = 2 (ad - bc) / [(a + c)(c + d) + (a +b)(b + d)]$$

El HSS mide la mejora fraccionaria de la previsión sobre la previsión del estándar. Como la mayoría de los scores de habilidad, se normaliza el rango total de posible mejora con respecto a la norma, lo que significa que los scores de habilidad Heidke con seguridad se puede comparar en diferentes conjuntos de datos. El rango de la HSS es $-\infty$ a 1. Los valores negativos indican que el pronóstico es mejor oportunidad, 0 significa que no hay habilidad, y una previsión perfecta obtiene un HSS de 1.

El HSS es un score popular, en parte porque es relativamente fácil de calcular y quizás también porque la previsión del estándar, el azar, es relativamente fácil de superar. Otras puntuaciones estándar son posibles, tales como la persistencia o la climatología, pero eso requiere de información adicional para calcular, en forma de una tabla de contingencia separada.

4. Resultados esperados y entregables

4.1 Modelo estadístico de pronóstico de escorrentía con 2 meses de antelación (Martín)

Los modelos estadísticos generados para este proyecto se hicieron con la colaboración y asesoría del Dr. Floris Van Ogtrop, experto en modelos estadísticos de la Universidad de Sydney, y tienen las siguientes características:

- están programados en el conocido lenguaje estadístico R,
- están basados en los modelos aditivos generalizados cuya teoría fue revisada con anterioridad,
- de todas las bases de datos disponibles toman aproximadamente el 50% de los datos para "construir" el modelo estadístico y el 50% restante para evaluar la calidad del pronóstico generado (*hindcast*),
- se evalúa la habilidad del pronóstico en la variable continua y discreta para el pronóstico a un mes determinado y posteriormente se evalúa toda la serie de tiempo de datos observados.

Las partes fundamentales del modelo estadístico son tres:

i) El ajuste de la ecuación por medio del modelo aditivo generalizado (gam) para una función continúa utilizando

```
mod2 <- gam(Flow~s(Flow.2, k = 3)+s(Rain.2, k = 3)+s(varind3.2, k = 3)
           +s(varind4.2, k = 3), data = data3, family = Gamma(link ="log"))
```

En donde la variable a pronosticar (predictando) es el flujo de escorrentía (Flow), para un mes en particular, en función de los predictores, dos meses antes, que son: el propio flujo de escorrentía (Flow.2), la precipitación (Rain.2), y una tercera (o más) variable(s) independiente(s) que generalmente está(n) asociada a las TSM y/o a una oscilación climatológica. La objeto "family" (en este caso Gamma) indica la descripción de la distribución de error y función de enlace que debe utilizarse en el modelo, se escoge una de entre 8 opciones como: bionomial, gausiana, inversa gausiana, poisson, causi, cuasibinomial, cuasipoisson y por supuesto Gamma. Se utilizó Gamma debido a que los datos continuos de precipitación y escorrentía siguen principalmente este tipo de función.

Las métricas calculadas para evaluar el desempeño de mod2 fueron varias, sin embargo en la tabla de resultados solo se muestran algunas de las más significativas como:

- la significancia aproximada de los términos suavizados de los predictores, en donde los rangos de los códigos son 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1;

- el porcentaje de desviación estándar explicado (*de*);

- el error medio absoluto (*mae*); y

- la eficiencia Nash-Sutcliffe (*nse*).

ii) El ajuste de la ecuación para una función discreta usando

```
modbin <- gam(bin~s(Flow.2, k = 3+s(Rain.2, k = 3)
              +s(varind3.2, k = 3)+s(varind4.2, k = 3)
              , data = data3, family = binomial)
```

En donde ahora la variable a pronosticar "bin" es una función binaria (discreta) que tiene únicamente los valores "1" si el flujo es mayor que la mediana del flujo total, o "0" si no lo es. Nuevamente esta función se describe en función de las variables predictoras anteriores, únicamente que ahora la familia de la distribución del error es binomial.

Para esta parte las métricas calculadas para evaluar el modelo fueron:

- la significancia aproximada de los términos suavizados de los predictores, en donde los rangos de los códigos son 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1;

- el porcentaje de desviación estándar explicado (*de*);

- la tabla de contingencia de los valores binarios de pronóstico (pro) y observados (obs) dados en la forma,

pro► / obs▼	0	1
0	6	5
1	3	6

donde el '1' significa que el valor observado o pronosticado está por arriba de la mediana del total de valores, y '0' significa lo opuesto. Para el ejemplo de arriba los valores concordantes fueron 12 y los discordantes 8, es decir una asertividad del 60%.

- y el índice de habilidad de Heidke (*hss*).

iii) El ajuste de la ecuación para una función continúa usando

```
mod3 <- gam(Flow~s(Flow.2, by = Month, k = 3)
            +s(Rain.2, by = Month, k = 3) +s(NINO12.2, by = Month, k = 3)
            +s(AO.2, by = Month, k = 3), data = data6, family = Gamma(link ="log"))
```

A diferencia del caso mod2, ahora el predictando, 'Flow', se pronostica para todos los datos mensuales disponibles.

En esta parte se volvió a evaluar con el mismo conjunto de métricas que en (i).

Los valores del análisis de resultados para estos tres ajustes se muestran en la **Fig. 4.1**. Se hizo un análisis individual de todas las variables climáticas enunciadas y explicadas en la figura. De acuerdo al análisis de evaluación, los resultados más sobresalientes para pronosticar el flujo continuo del mes de septiembre con dos meses de antelación fueron dados por el modelo,

```
mod2 <- gam(Flow~s(Flow.2, k = 3)+s(Rain.2, k = 3)+s(NINO12.2, k = 3)
            +s(AO.2, k = 3), data = data3, family = Gamma(link ="log"))
```

Es decir la combinación aditiva del flujo, lluvia, TSM de la región Niño 1+2, y la Oscilación del Ártico. Es de notar la contribución tan importante de esta última variable ya que no hay nada documentado en la literatura sobre la influencia que tiene la Oscilación del Ártico sobre la zona del Monzón de Norteamérica. Los resultados del porque este caso fue considerado como el mejor pronóstico es debido a los valores bajo de *MAE* y el valor entre 0 y 1 del *nse*.

Para el caso del pronóstico de flujo de forma discreta el modelo que mejor funcionó fue:

```
modbin <- gam(bin~s(Flow.2, k = 3+s(Rain.2, k = 3)
              +s(NINO4.2, k = 3)+s(AO.2, k = 3)
              , data = data3, family = binomial)
```

Es decir la combinación aditiva del flujo, lluvia, TSM de la región Niño 4, y nuevamente la Oscilación del Ártico. El éxito de este caso esta mostrado por los valores de la tabla de contingencia con un total de 15 aciertos de 20 posibles (75% de efectividad) y el valor alto del índice de habilidad de Heidke (hss).

pro► obs▼	0	1
0	11	0
1	5	4

A continuación se mostrarán y describirán los productos más relevantes del mejor modelo construido para pronosticar el flujo continuo del mes de septiembre con dos meses de antelación, esto es con la combinación: Flow.2+Rain.2+NINO12.2+AO.2.

	mod2.s1	mod2.s2	mod2.s3	mod2.de	mod2.mae	mod2.nse	modb.s1	modb.s2	modb.s3	modb.de	modb.tc	modb.hss	mod3.s1	mod3.s2	mod3.s3	mod3.de	mod3.mae	mod3.nse
NINO12.2			*	49.8	87.156	-0.543			*	74.9	0506-0603	0.117	**	***	***	35.5	98.145	-0.435
NINO3.2				28.3	91.228	-2.53			*	62	0703-0406	-0.031	***	***	*	27.9	97.832	-0.533
NINO4.2			.	30.5	83.55	-3.342			*	52.9	0904-0205	0.27	***	***	***	31.7	101.695	-0.346
NINO34.2				24.7	86.639	-3.321	*		*	65.2	0803-0306	-0.062	**	***	***	30.3	90.763	-0.666
IOD.2				29	81.746	-3.003				39.8	0604-0505	-0.01	***	***		26.5	111.957	-0.355
PDO.2				24.9	84.507	-5.847				35.9	1002-0107	0.14	***	***	**	29.2	107.272	-0.296
AMO.2				21.2	89.665	-6.219				20.9	0706-0403	0.3	***	***		26.7	106.867	-0.416
NAO.2				32.2	85.338	-2.247				27	0606-0503	0.208	***	***		26.6	111.967	-0.361
AO.2			*	56.3	76.837	-1.287				44.4	0904-0205	0.271	***	***	***	30.4	116.959	-0.325
NINO12.2/AO.2			**/**	72.9	64.603	0.194	.	*	*/_	79	0606-0503	0.208	**	***	***/***	38	102.887	-0.379
NINO4.2/AO.2			_/*	55.6	75.608	-2.038	***	***	***/***	100	1104-0005	0.468	***	***	***/***	36	103.017	-0.317
NINO12.2/NINO4.2			*/_	49.4	87.241	-0.593	***	***	***/***	99.9	0506-0603	0.118	**	**	***/***	38.5	94.393	-0.532
NINO12.2/NINO4.2/AO.2			**/_/**	72.5	67.166	0.073	***	***	***/***/***	99.8	0606-0503	0.208	**	**	***/***/**	41.3	99.211	-0.447

Fig. 4.1 Resultados de validación para el modelo estadístico de pronóstico de escorrentía de septiembre con 2 meses de antelación (esto es, julio).

Notación: NINO12.2 – SST de la zona Niño 1+2 con desfasamiento de dos meses (‘.2’); NINO3.2 – SST zona Niño 3; NINO4.2 – SST zona Niño 4; NINO34.2 – SST zona Niño 3.4; IOD.2 – Índice del Dipolo del Océano Índico; PDO.2 – Oscilación Decadal del Pacífico; AMO.2 – Oscilación Multidecadal del Atlántico; NAO.2 – Oscilación del Atlántico Norte; AO.2 – Oscilación del Ártico; NINO12.2/AO.2 – Introduce esas dos variables en la formulación del modelo, etc. Por otro lado mod2.s1, mod2.s2, mod2.s3, indica que tan significativo es el valor de suavizamiento del flujo continuo pronosticado como función del flujo, lluvia y variable(s) con dos meses de antelación de acuerdo a los rangos de los códigos son 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1. mod2.mae es el error absoluto medio del flujo pronosticado y mod2.nse es la eficiencia Nash-Sutcliffe. Esas mismas métricas se aplican a mod3 (pronóstico del flujo para toda la serie de tiempo con dos meses de antelación). Para modb.s1, modb.s2, modb.s3 y modb.de se aplica la misma interpretación que para mod2 y mod3. modb.tc es la tabla de contingencia del flujo pronosticado en forma binaria (ver más detalle en el texto) donde los primeros 4 dígitos se interpretan de la siguiente forma: por ejemplo 1104-0005, el ‘11’ indica el número de veces que tanto lo observado como lo pronosticado coinciden en que el flujo fue menor que la mediana, el ‘04’ indica el número que tanto que tanto lo observado como lo pronosticado coinciden en que el flujo fue mayor que la mediana, de esta forma el total de aciertos del pronóstico es ‘15’; después el ‘00’ indica que no hubo casos en que lo observado haya estado por debajo de la mediana y se haya pronosticado lo contrario, y ‘05’ indica el número de veces en que lo observado estuvo por arriba de la mediana y se pronosticó lo contrario, de esta forma el número total de fallas fue de ‘5’, con una efectividad del 75%. El modb.hss es el valor de habilidad de Heidke. Los valores resaltados en amarillo y naranja fueron los resultados con mejor efectividad para el caso continuo del pronóstico del flujo y del caso binario respectivamente.

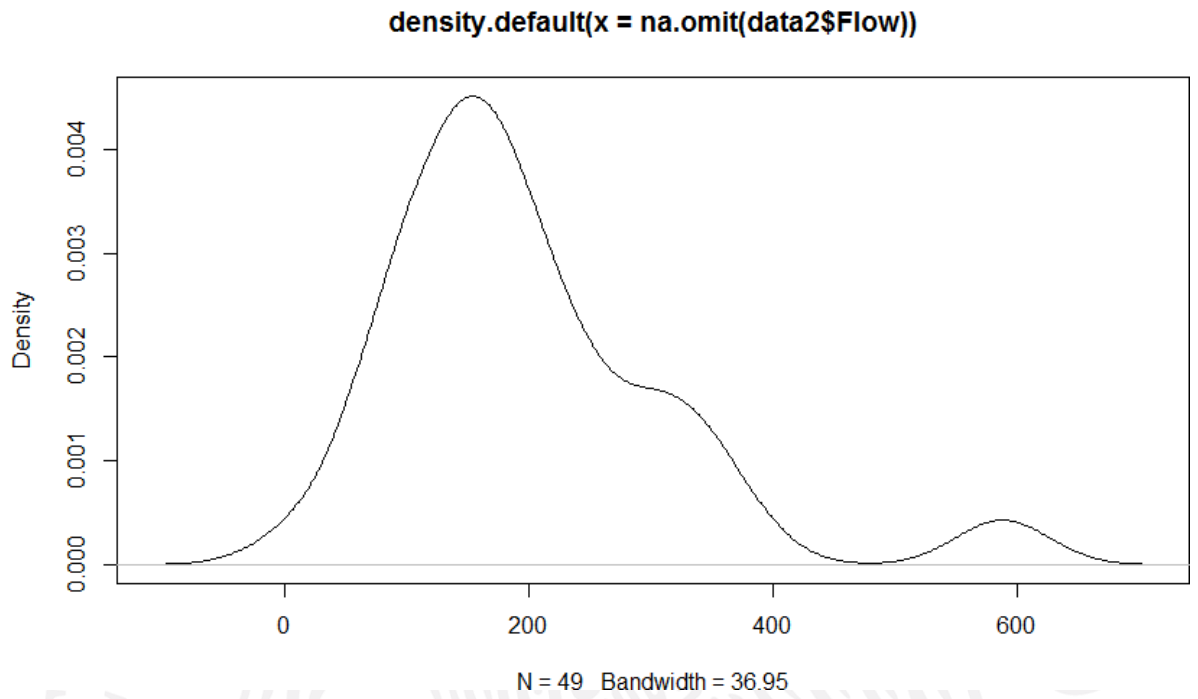


Fig. 4.2 Función de densidad del flujo medido en Huites (en Hm3) para el mes de septiembre en el período 1959-2007.

La función de densidad del flujo (Flow) medido en la estación hidrométrica Huites para el mes de septiembre se muestra en la **Fig. 4.2**. N denota el número de meses 'septiembre' para toda la base de datos (en este caso 49 años de 1959 a 2007), el eje de la abscisa es el valor del flujo medido en Hm^3. De la misma forma, el histograma de frecuencias correspondiente se muestra en la **Fig. 4.3**.

El grado de significancia de los términos de suavizamiento de cuada una de las cuatro variables que constituyen el modelo se muestran en la **Fig. 4.4**.

Los resultados de los resultados del modelo gam mod2 construido aquí se muestran en la **Fig. 4.5**. Un buen resultado se muestra cuando la desviación de los residuales sigue la línea recta como en la gráfica de arriba a la izquierda. También un buen indicador del ajuste es que los valores de los residuales están distribuidos homogéneamente (sin sesgo) en la gráfica de arriba a la derecha.

Los resultados de los pronósticos del flujo y su comparación con lo observado se muestran en la **Fig. 4.6**. En general, se nota una muy buena correspondencia entre lo observado y lo pronosticado.

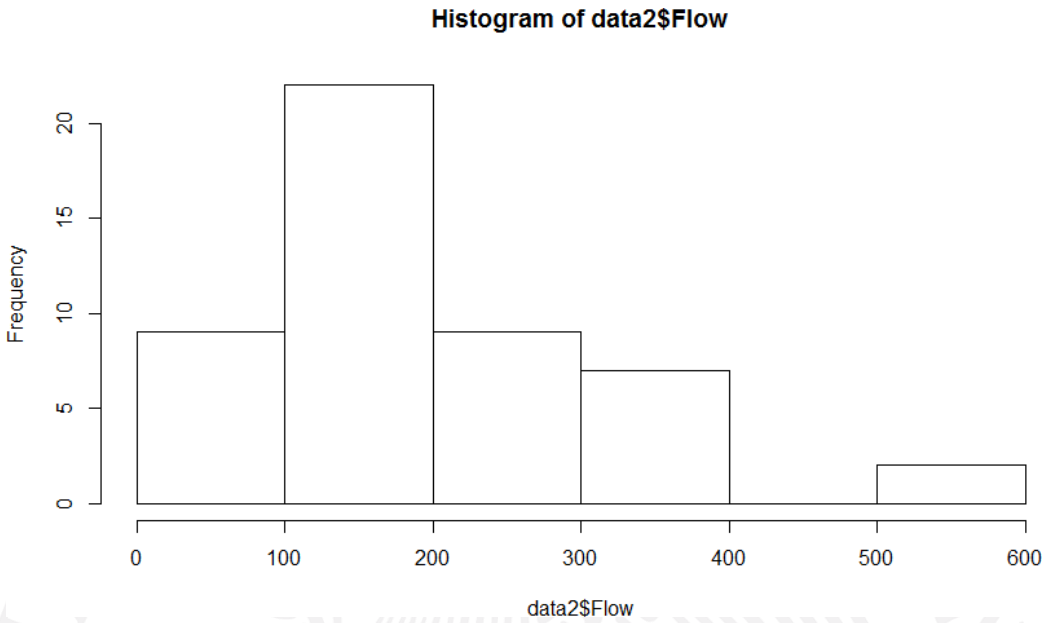


Fig. 4.3 Histograma de frecuencias del flujo medido en Huites (en Hm3) para el mes de septiembre en el período 1959-2007.

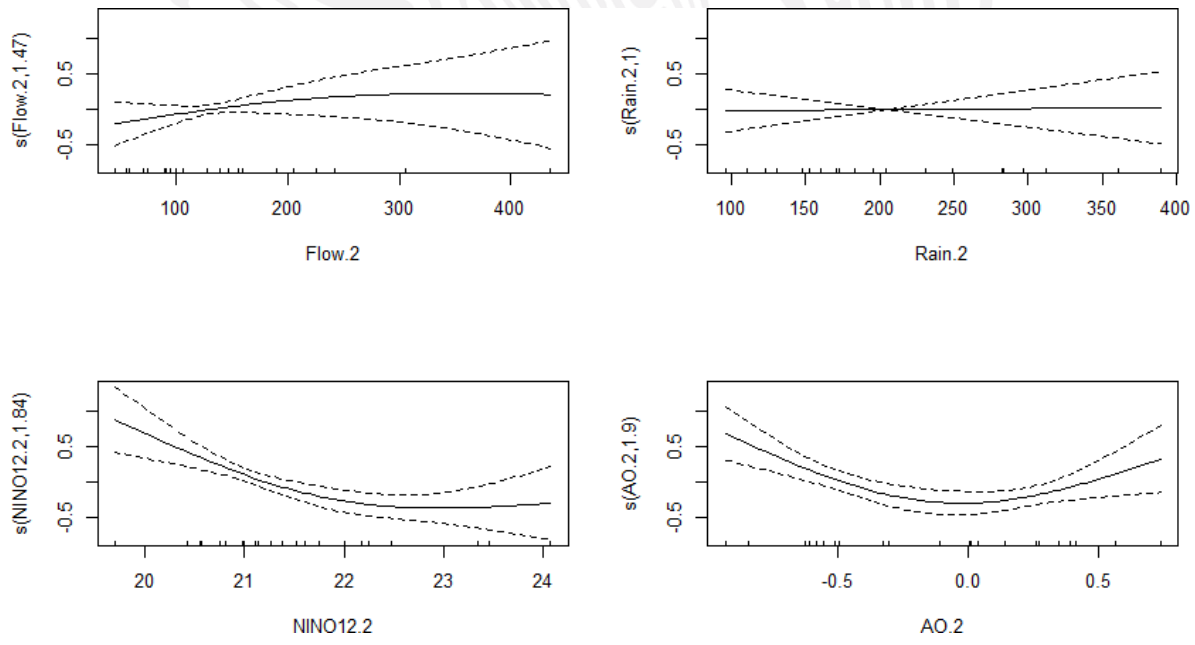


Fig. 4.4 Valor de significancia de los términos de suavizamiento del Flujo (arriba izq.), Lluvia (arriba der.), TSM de la región Niño1+2 (abajo izq.) e Índice de la oscilación del Ártico (abajo der.).

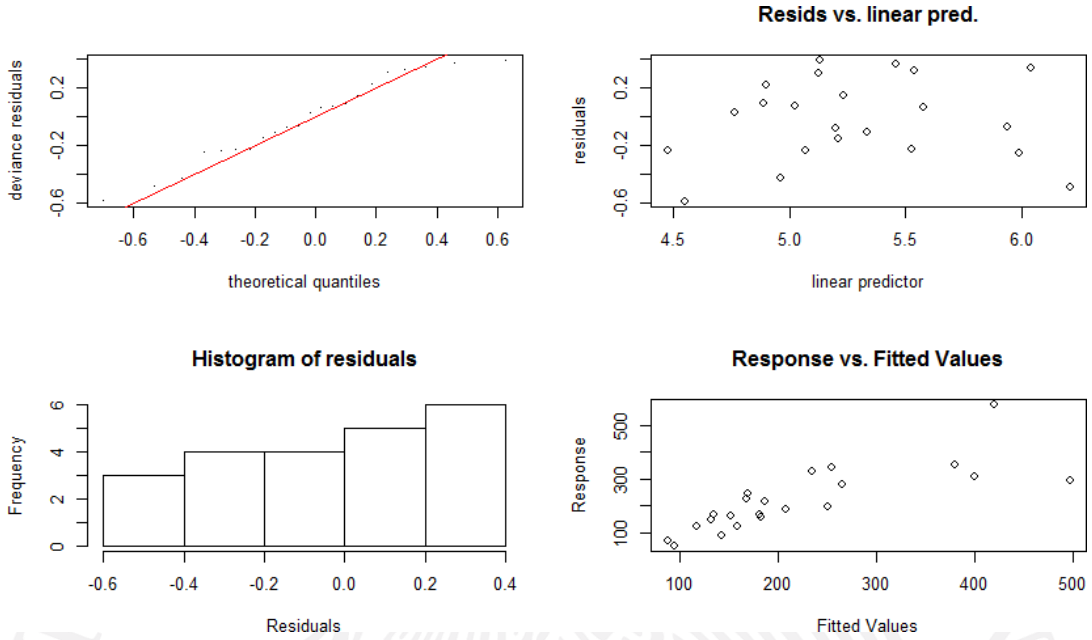


Fig. 4.5 Resultados de los residuales del modelo gam mod2. Se muestran la desviación de los residuales vs cuantiles teóricos (arriba izq.), los residuales vs el predictor lineal (arriba der.), el histograma de frecuencia de residuales (abajo izq.) y la respuesta vs valores ajustados (abajo der.).

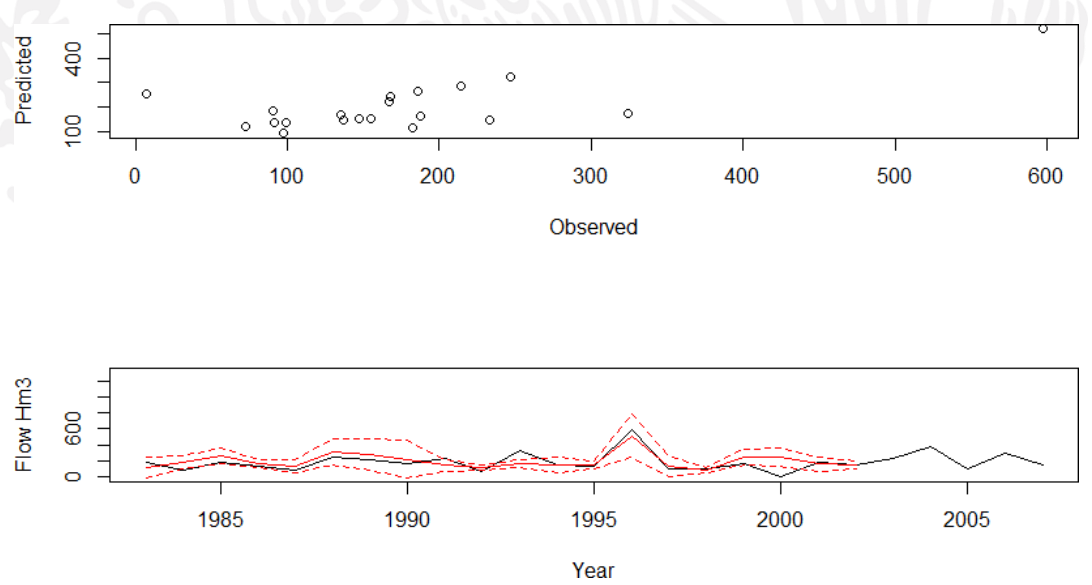


Fig. 4.6 Gráfica de dispersión de los valores del flujo (en Hm3) pronosticados contra los observados (arriba) y serie de tiempo de los valores observados (línea negra sólida) y los valores pronosticados (línea roja sólida) con sus valores de incertidumbre asociados (líneas rojas punteadas).

4.2 Modelo estadístico de pronóstico del total de escorrentía para los períodos, Junio-Septiembre y Septiembre-Diciembre, con varios meses de antelación.

Los resultados de este entregable fueron proporcionados por el Dr. Floris Van Ogtrop, colega y colaborador de la Universidad de Sídney. El reporte enuncia los resultados del pronóstico para escurrimiento total acumulado del período Junio-Septiembre realizado en Abril y del período Septiembre-Diciembre realizado en Julio.

Summary of results for forecasts – Mexico 2014

Floris van Ogtrop

The following report is a brief summary of initial findings for forecasting total inflows for the period June through September for the period 1961 to 2002. Initial investigation only looked at a one month lag i.e. regressing April SSTs with total June – September flow.

Three methods where considered, continuous (volumetric) forecasts, categorical forecasts and finally a simple SOI phase based continuous forecast. The results are presented in this order. The VGAM package was used for the first two exercises as this package could run the tercile forecasts.

Continuous forecasts

Figure 2 shows the continuous forecasts of total June to September and September to December inflows into Huites in cubic hectometres. The regression model regressed the total inflows with the average April – March SSTs and June – August SSTs. The following indices were used Nino1+2, Nino3, Nino3.4, Nino4, IOD, AMO, SWM, NAO, and AO. Some initial experimentation was done with including different indices and lagged Flow, Nino1+2, IOD, AMO and AO, in various combinations, seemed to be the best predictors of total flow. All were significant ($p<0.05$). Though the model selection could be more rigorous, for the sake of demonstration these indices were chosen. The models were trained on the first 70% (1961 – 1992) of the data and validated on the second 30% (1993 – 2002) of the data. The data was assumed to be normally distributed (i.e. **Fig. 4.7**).

The models were not able to fully capture the variability in the observed data (**Fig. 4.8**) and this is reflected in poor goodness of fit scores (**Table 4.1**). However, it may be possible to derive general “rules of thumb” from **Fig. 4.9**. For example decreasing Nino 1+2 seems to have a positive effect on total inflows. Similarly the further the IOD value is from 0 the less total inflows are. Furthermore, an increasing AO and value greater than 1 may be associated with increasing inflows. These relationships could be further investigated and importantly tested to see whether they are robust in time. In addition, the bottom right plot in **Fig. 4.8** shows that the model is not constrained to positive values. This may suggest that the gamma distribution should be used instead of the lognormal distribution as this will constrain the model to positive values. However, this was not further investigated.

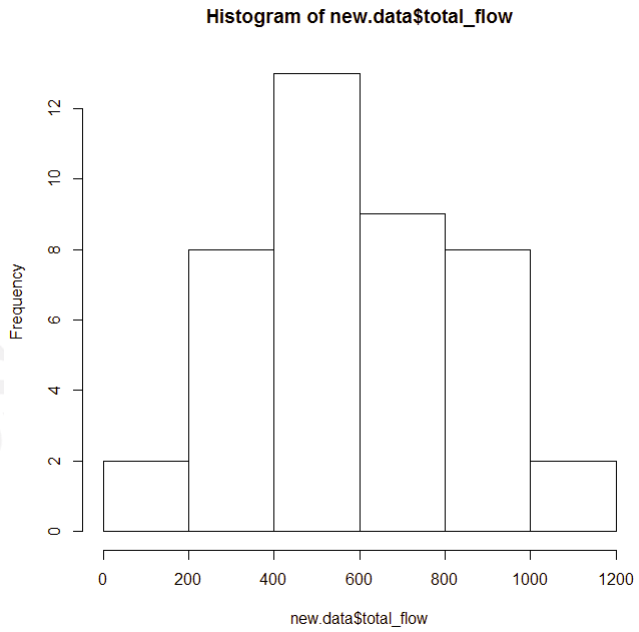


Fig. 4.7 Distribution of Total June - September inflows.

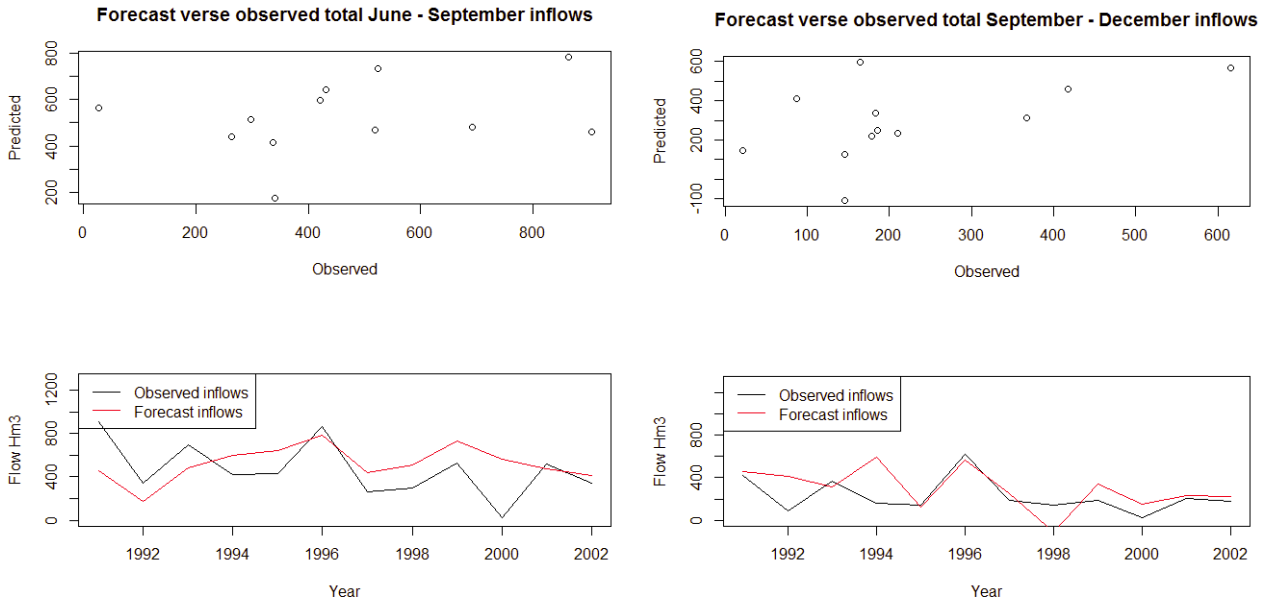


Fig. 4.8 Observed versus Forecast total June – September inflows scatterplot (top left) and timeseries plot (bottom left) and observed versus Forecast total September – December inflows scatterplot (top right) and timeseries plot (bottom right).

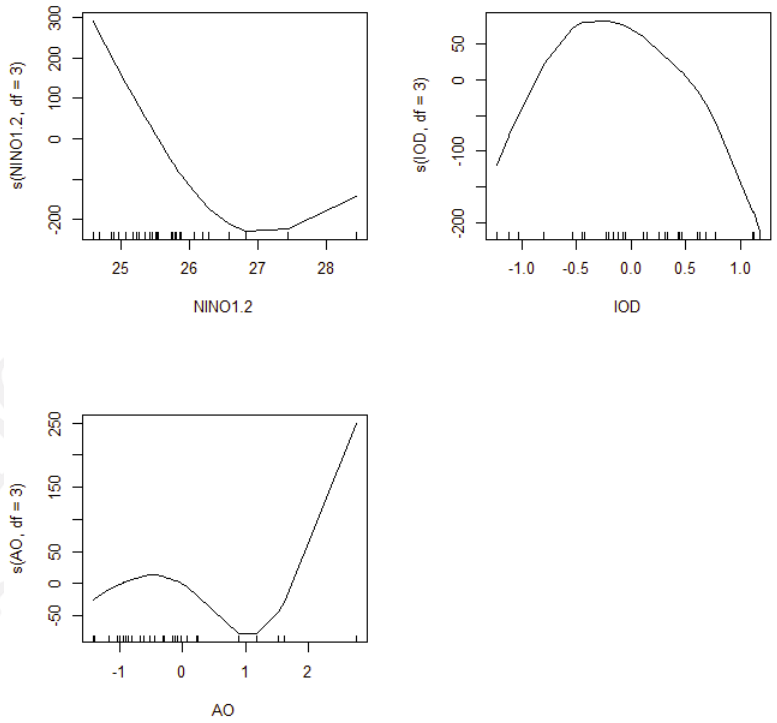


Fig. 4.9 Relationships between SST indices and total inflows

Categorical Forecast

The categorical forecasts were again completed using the vgam function in the VGAM package (Yee 2014; Yee and Wild 1996). The cumulative family was used as the appropriate model for categorical data. The final results show reasonable skill in predicting the terciles as low (0-25 percentile), medium (25-75 percentile) and high (75 – 100 percentile) (**Fig. 4.10**). Similarly, **Table 4.2** and **Table 4.3** show the forecast successes and failures. The numbers highlighted in red in **Table 4.2** and **Table 4.3** show where the forecasts predicted more inflows than occurred and are by default the most worrisome as they may lead to overconfidence in decisions made such as area planted, crop type selected, or volumes of water made available.

SOI Phases

The final forecasting tool that was explored is looking at the SOI phases. Initially simple box plots have been used to look at whether there is a difference between the distributions using boxplots (**Fig. 4.11**) of total June – September inflows for the different phases in a preceding month (April in this example). In this example the data set was split in half so the training set was from 1961 – 1981 and the validation set 1982 – 2002. For this method to be useful, it would be expected that the relationship is similar between the two sets. From **Fig. 4.11**, it is clear that this is relationship is not consistent for this lagged relationship. Further exploration may find situations where this type of forecast may be useful. Furthermore, given the limited available data, it may be some years before there is enough data to establish these relationships.

Table 4.1 Goodness of fit measures for continuous forecasts descriptions of the measures can be found at http://127.0.0.1:22457/library/hydroGOF/html/gof.html

Goodness of fit measure	Jun-Sept	Sept-Dec
ME	-54.41	-68.91
MAE	212.61	131.51
MSE	64121.94	131.51
RMSE	253.22	184.62
NRMSE %	158.1	93
PBIAS %	-10.4	-23.3
RSR	1.58	0.93
rSD	1.58	0.82
NSE	-1.73	0.06
mNSE	-0.81	0.13
rNSE	-2.52	-0.54
d	0.5	0.71
md	0.34	0.57
rd	0.36	0.53
cp	-0.36	0.58
r	0.28	0.52
R2	0.08	0.28
bR2	0.07	0.18
KGE	0.07	0.44
VE	0.59	0.56

Table 4.2 Table of forecast successes and failures June – September forecast

		Forecast		
		Low	Medium	High
Observed	Low	**3**	**3**	0
	Medium	0	**3**	**1**
	High	1	0	**1**

Table 4.3 Table of forecast successes and failures September – December forecast

		Forecast		
		Low	Medium	High
Observed	Low	**4**	**3**	**3**
	Medium	0	**2**	**1**
	High	1	0	**1**

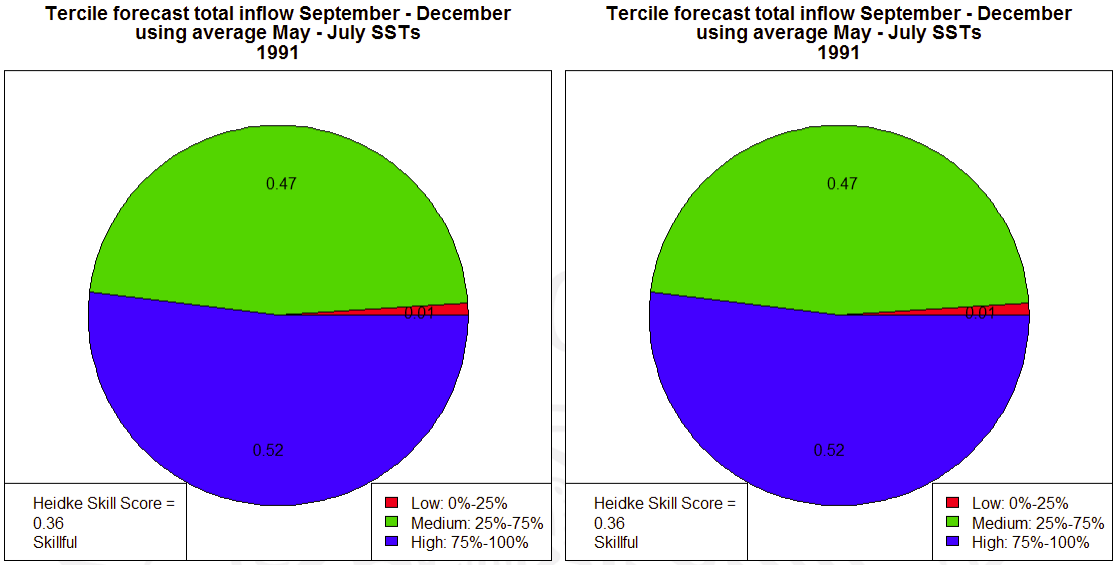


Fig. 4.10 Tercile forecasts of total inflow for June to September based on average February - April SSTs (Nino 1+2, IOD and AO) (Left) and tercile forecasts of total inflow for September to December based on Average May - July SSTs (lagged Flow, Nino 1+2, IOD and AMO).

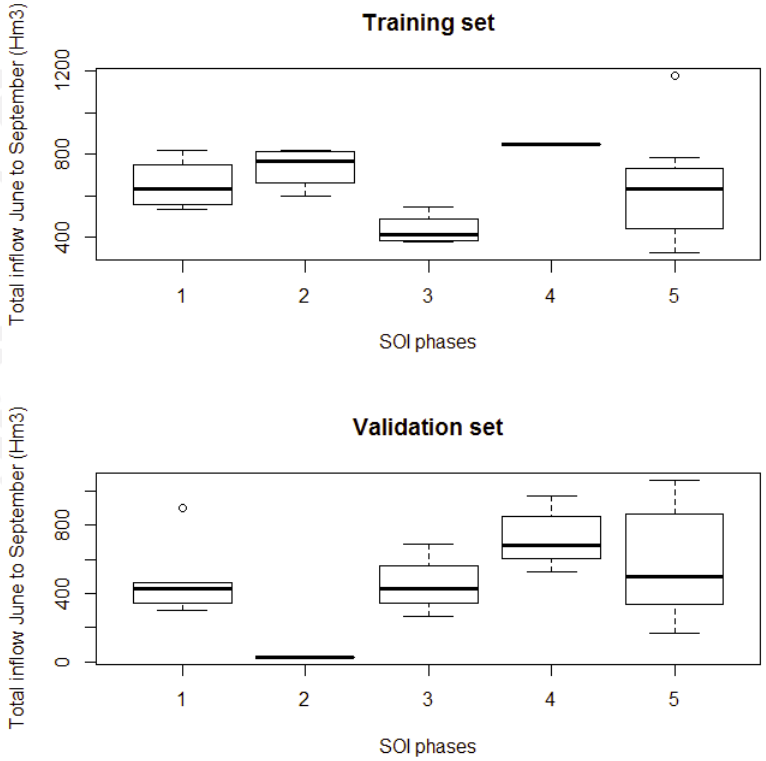


Fig. 4.11 Boxplots of distributions of June - September total inflows based on April SOI phases

4.3 Generación de productos agroclimatológicos que sean de mayor utilidad a los agricultores de zonas agrícolas bajo riego a partir de los pronósticos estacionales que ya se generan a nivel nacional

Los nuevos retos de la agricultura demandan un uso óptimo de los insumos que utiliza bajo un esquema de variabilidad climática, alta competitividad, incremento en costos de producción, volatilidad de precios y riesgo ambiental. La recurrencia de sequías y la competencia por el agua son la principal amenaza al desarrollo y convivencia armónica de las zonas agrícolas, principalmente para las zonas riego mayormente localizadas en regiones áridas y semiáridas donde el riego es indispensable para obtener rendimientos comerciales. Bajo condiciones de baja disponibilidad y alta competencia por agua, se requiere de acciones y herramientas para utilizar óptimamente los recursos hídricos disponibles sin una reducción significativa en las superficies y rendimientos convencionales.

En general, la mayoría de las tomas de decisiones agrícolas se realiza con un alto nivel de incertidumbre debido a la naturaleza de la información disponible: i) incertidumbre debido al limitado conocimiento del estado actual del sistema planta-ambiente, ii) incertidumbre debido a un conocimiento parcial de los sistemas biológicos y físicos asociados al sistema de producción agrícola, y por último, iii) incertidumbre debido a los procesos aleatorios inherente a los sistemas biológicos y físicos que se presentan no solo a nivel parcela sino también a nivel de la cuenca que abastece las fuentes de aprovechamiento de las zonas de riego. Ante la incertidumbre en la información requerida para planear sus actividades de un año agrícola, los agricultores minimizan el riesgo simplificando sus sistemas productivos de forma muy conservadora. Por ejemplo, ante una baja dotación de agua parcelaria, los agricultores se adaptan a las restricciones en la dotación asignada. Diferentes alternativas eligen los agricultores, como reporta Ojeda Bustamante (2002), al reducirse su dotación normal usualmente se recurre a reducir la superficie cultivada por productor, a seleccionar cultivos de baja demanda de agua, a extraer aguas de fuentes alternas como agua subterránea, y a establecer programas de emergencia para optimar el uso del agua.

Ante este panorama, la herramienta del pronóstico estacional apoyado en las tecnologías de información es un medio para proporcionar información con oportunidad y analizar escenarios de planes de cultivos con la finalidad de reducir la incertidumbre en la toma de decisiones agrícolas antes y durante la ejecución de un año agrícola. Rebgetz *et al* (2005) presentan los beneficios potenciales del pronóstico estacional de variables hidroclimáticas de interés para la agricultura de riego: dotación de agua, precipitación y evapotranspiración. Existe el consenso que el pronóstico climático estacional es de enorme utilidad para ajustar decisiones agrícolas críticas, sin embargo, como comentan Hansen *et al.* (2006), hay una laguna entre la información que suministran los centros de predicción climática y la necesidad de los agricultores y tomadores de decisiones agrícolas.

La herramienta del pronóstico estacional es de utilidad para conocer dos variables básicas que definen muchos de los procesos agrícolas que definen la producción y eficiencia de los sistemas agrícolas: i. precipitación y ii. temperatura. El conocimiento anticipado, con un nivel probabilístico, de la precipitación permitiría estimar con mayor precisión las demandas de riego de los cultivos. El desarrollo de los cultivos se asocia a la temperatura. Altas temperaturas acortan los ciclos y bajas

temperaturas los alargan, por lo que el conocimiento de los cambios en las temperaturas esperadas puede ser de utilidad para conocer con anticipación cambios esperados en la fenología de los cultivos.

En cuanto a variables secundarias, la herramienta del pronóstico estacional es de utilidad para conocer dos variables que son de importancia para planear un año agrícola: i. evapotranspiración y ii) escurrimiento. La primera depende de la temperatura y humedad ambiental, la radiación solar y la velocidad del viento y la segunda de la precipitación.

Una de las aplicaciones comerciales de pronóstico estacional de escurrimientos en diferentes puntos de interés de una cuenca se usa en el estado de NWSW al oeste de Australia, un servicio suministrado por el Bureau of Meteorology (BOM) de Australia. La **Fig. 4.12** presenta el pronóstico estacional de enero a marzo de 2014 en términos probabilísticos en diferentes cuencas de interés. En la **Fig. 4.13** se presentan los escurrimientos mensuales históricos como referencia. Este tipo de pronóstico estacional es de interés para México para los distritos de riego abastecidos por presas de almacenamiento que suministran el 75% de la superficie regadas de los distritos de riego del país, la mayoría localizados en zonas áridas y semiáridas.

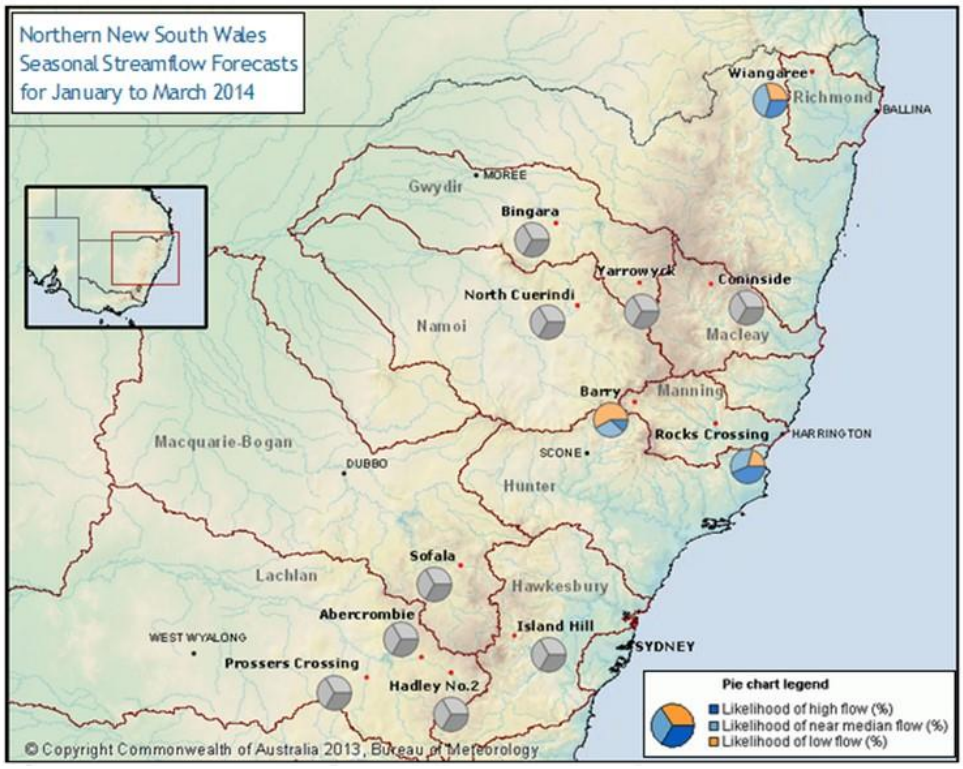


Fig. 4.12 *Pantalla de consulta del pronóstico estacional de escurrimientos*

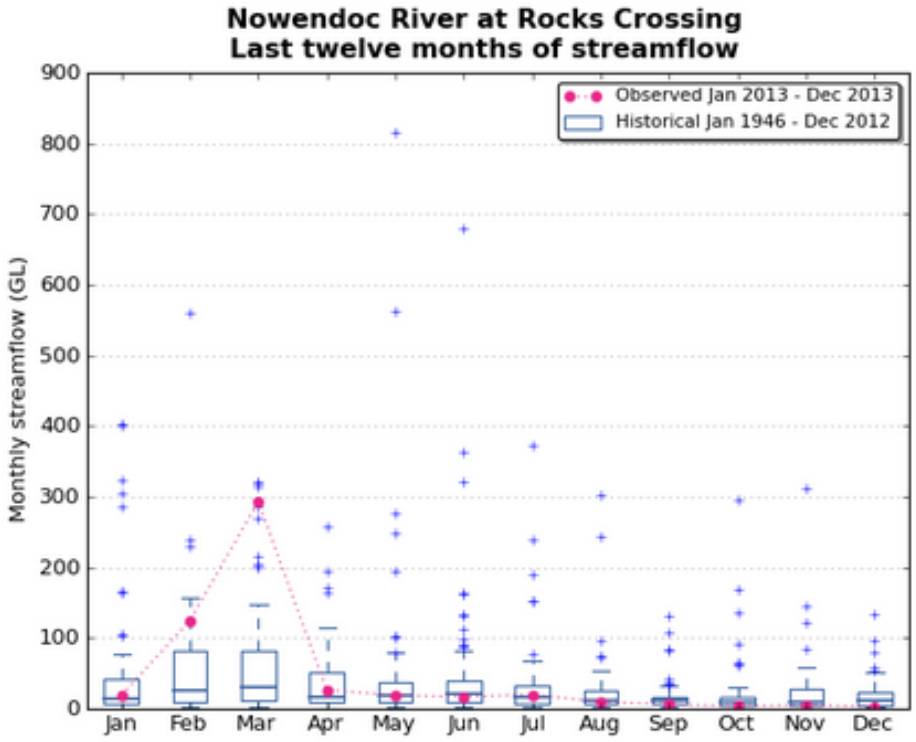


Fig. 4.13 *Consulta de escurrimientos históricos de una cuenca*

Cuando el resultado del pronóstico estacional se acopla a modelos biológicos, es posible generar productos de mayor interés para la agricultura. La **Fig. 4.14** presenta el producto de un sistema de pronostico operacional para la producción de trigo a nivel municipal para el mes de junio de 2002 para Australia (Stone y Meinke, 2005). El método usado (Potgieter et al. 2002) emplea un modelo empírico de pronostico climático estacional (Stone et al. 1996) conectado a un modelo agroclimático hibrido, de donde un índice es derivado de la grado de estrés hídrico relativo a la disponibilidad del agua para la planta (Hammer et al. 1996). La leyenda se refiere a la probabilidad de exceder el rendimiento medio de largo plazo, relatico para cada valor municipal.

Sin duda el servicio agroclimático basado en la aplicación del pronóstico estacional de variables de interés agrícola como temperatura, precipitación, escurrimiento o evapotranspiración es una necesidad de la agricultura mexicana como instrumento de gestión de riesgo y para planear mejor las actividades agrícolas involucradas para la toma de decisiones.

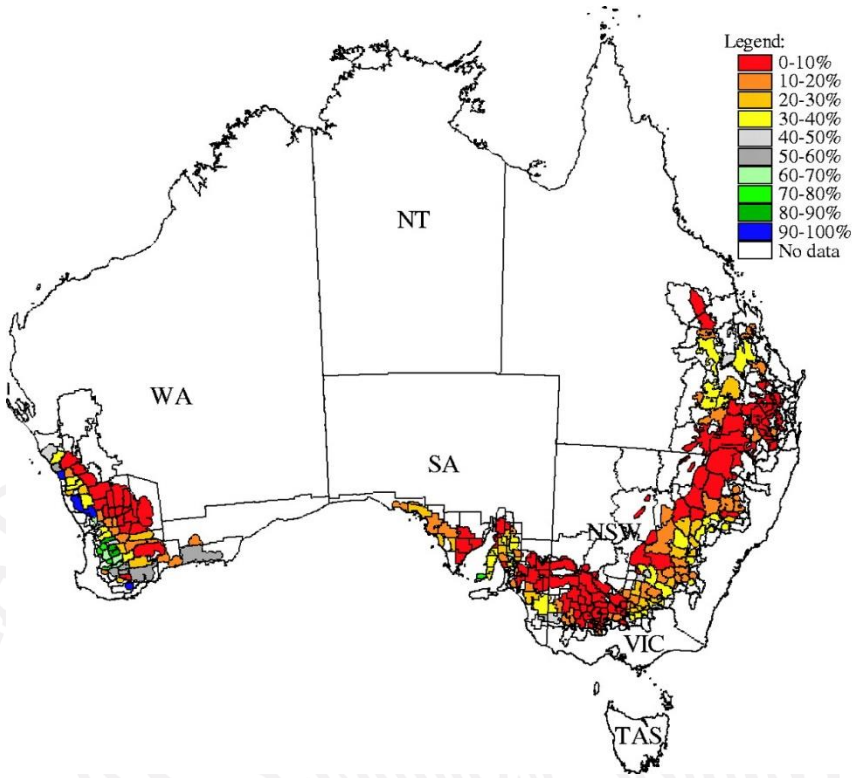


Fig. 4.14 Ejemplo de la aplicación de un modelo de pronóstico estacional operacional para evaluar el riesgo del trigo

4.4 Diseño de un proyecto para aplicación práctica de los conocimientos desarrollados en una zona de riego de importancia para México.

Como resultado de la Mesa de Trabajo en el área de Agroclimatología durante la Misión Australia en México 2014, el día 25 de noviembre de 2014 se generó una propuesta de trabajo en colaboración con todos los integrantes de este proyecto, el Dr. Floris Van Ogtrop (U. Sydney) y el Dr. Luis Rendón Pimentel, Gerente de Distritos de Riego de la CONAGUA.

COOPERACIÓN MÉXICO- AUSTRALIA EN MATERIA DE RECURSOS HIDRÁULICOS MESAS DE TRABAJO
PROPUESTAS DE PROYECTOS DE COOPERACION TÉCNICA INTERNACIONAL

1. Título del Proyecto

Análisis del impacto del cambio climático y variabilidad climática en el recurso hídrico para la planeación de riego de la cuenca del Rio Yaqui.

2. Antecedentes

El Río Yaqui es uno de los distritos de riego más importantes de México y tiene muchas similitudes con cuencas de ríos grandes en Australia. El área de riego es alrededor de 300,000 Ha. Además tiene disponible bases de datos climáticas y agronómicas de períodos largos que pudieran tener el potencial de probar ya señales de cambio climático. De acuerdo a los a registros históricos, ha habido sequías importantes en esa región y tiene ya problemas de salinidad. También existen problemas socio económicos importantes en cuanto a la distribución de agua con ciertos usuarios y habitantes de la región. También los modelos climáticos globales indican que la región es altamente vulnerable a cambio climático.

3. Contexto

- Es una cuenca de río vulnerable.
- Hay muy pocos estudios que analizan los impactos de cambio climático y variabilidad en cuencas hidrológicas de México.

4. Instituciones involucradas

- IMTA
- Universidad de Sídney
- Conagua
- UNISON
- ITSON
- INIFAB
- Asociación de Usuarios del Agua

5. Objetivo General del Proyecto

- Análisis estadístico de los datos históricos para verificar si ya existen señales de cambio climático en la cuenca.
- Desarrollar un modelo hidrológico para la cuenca (acoplado a un modelo atmosférico regional) para estimar el impacto de proyecciones climáticas futuras de la temperatura, la lluvia y el caudal.

6. Objetivo Específico del Proyecto

- Recopilar datos hidrológicos para la cuenca del río
- Utilizar el conjunto de datos de alta calidad del SMN para realizar análisis de homogenización para detectar si las señales del cambio climático ya están presentes en las variables relevantes para la agricultura y la gestión del agua
- Proyecciones del CMIP5 ya reducidas de escala dinámicamente utilizando el modelo regional (WRFC y RegCM4)
- Ejecutar los modelos hidrológicos (SWAT y SOURCE) con salida del modelo regional.

7. Resultados

- Base de datos de alta calidad para la región (precipitación, temperatura, caudal, velocidad del viento, etc.)
- Un modelo hidrológico calibrado para la cuenca.
- Respuesta a la pregunta ¿hay actualmente pruebas de alguna señal de cambio climático en la cuenca?
- Identificar los posibles cambios en extremos climáticos futuros tales como sequías, flujos bajos, salinidad, olas de calor, etc.
- Escribir trabajos de colaboración en investigación y nuevas propuestas con los miembros del equipo
- Fortalecer las relaciones de colaboración entre científicos mexicanos y australianos

8. Actividades

- Revisión de literatura.
- Las visitas de campo para la recogida de datos y la consulta a la comunidad.
- Las visitas de campo para validación del modelo.
- Talleres de capacitación y videos de información para miembros de la asociación de usuarios del agua y los administradores locales de agua.
- Intercambio de estudiantes.
- Análisis de datos y elaboración de modelos.
- Preparar informes y trabajos de investigación relacionados a la cuenca.

9. Modalidades de colaboración

Visitas técnicas Australia-México.

10. Duración del Proyecto

3 años.

11. Presupuesto estimado por rubros

Concepto.	Cantidad (Millones USD)
Equipo de cómputo robusto para llevar a cabo la modelación dinámica regionalizada acoplada con modelos hidrológicos que es fundamental para el proyecto.	1.0
Adquisición de software y de espacio de almacenamiento informático porque las salidas que producen esos programas de cómputo son muy grandes.	0.2
Visitas de campo a la cuenca del Río Yaqui de un equipo de alrededor de al menos 12 investigadores de las 7 instituciones relacionadas al proyecto.	0.2
Estancias de investigación mutuas entre investigadores de Australia y México para	0.3

intercambio de conocimientos y resultados del proyecto.	
Gastos diversos de las otras instituciones envueltas en el proyecto además de IMTA, CONAGUA y U. de Sydney (esto es para UNISON, ITSON, INIFAP, Asociación de Usuarios del Agua).	0.3
Gran Total por 3 Años	2.0

12. Fuentes de financiamiento (potenciales)

- CONAGUA
- CONACYT
- AusID
- Banco Interamericano de Desarrollo
- Banco Mundial

5. Conclusiones

El presente proyecto ha mostrado el potencial de los modelos aditivos generalizados como herramientas de pronóstico estacional para predecir el flujo (escurrimiento) hasta con dos (o más) meses de anticipación de cuencas con poca influencia antropogénica como es la de Huites estudiada aquí.

De acuerdo a los resultados encontrados aquí, el factor climatológico de mayor relevancia para la zona fue la Oscilación del Ártico (AO) debido a que junto con la combinación de la influencia de la zona Niño 1+2 y de la zona Niño 4 se predijo con un buen grado de certidumbre los flujos continuos (caso mod2) y flujos divididos por categorías (modbin) respectivamente. En el caso del flujo continuo se alcanzó una excelente eficiencia Nash-Sutcliffe de 0.2, y para el caso del flujo por categorías se alcanzó una efectividad de 75% en la tabla de contingencia y un índice de habilidad de Heidke alto (0.47) de acuerdo a los resultados de la **Fig. 4.1**.

Por otro lado, los resultados encontrados por nuestro colega australiano, el Dr. Van Ogtrop, con una variación del modelo estadístico usado en este proyecto (los modelos aditivos generalizados vectoriales), muestran señales claras de predecibilidad para el flujo total acumulado en los períodos Junio-Septiembre y Septiembre-Diciembre, ambos períodos muy importantes para la planeación hídrica y agrícola de la zona. Los modelos vectoriales que mejores resultados arrojaron fueron: i) una combinación de la señal de Niño 1+2, IOD y AO promediada de Febrero–Abril para el pronóstico en terciles del Flujo Acumulado de Junio–Septiembre; y ii) una combinación de la señal del Flujo, Niño 1+2, IOD y AMO promediada de Mayo–Julio para el pronóstico en terciles del Flujo Acumulado de Septiembre–Diciembre.

En cuanto a generar nuevos productos agroclimatológicos se mencionó que el pronóstico estacional, con un nivel probabilístico, de la precipitación permitiría estimar con mayor precisión las demandas de riego de los cultivos. Por otro lado, la temperatura es muy importante para el desarrollo de los cultivos, debido a que altas temperaturas acortan los ciclos y bajas temperaturas los alargan, por lo que el conocimiento de los cambios en las temperaturas esperadas puede ser de utilidad para conocer con anticipación cambios esperados en la fenología de los cultivos. Adicionalmente, el pronóstico estacional es muy importante para estimar el comportamiento futuro de variables que inciden en la planeación del año agrícola como la evapotranspiración y el escurrimiento. Se ponen como ejemplos de aplicación un par de productos concretos generados por el BOM de Australia.

Finalmente, como resultado de la Mesa de Trabajo en el área de Agroclimatología durante la Misión Australia en México en noviembre de 2014, se generó una propuesta de trabajo en colaboración con todos los integrantes de este proyecto, el Dr. Floris Van Ogtrop (U. Sydney) y el Dr. Luis Rendón Pimentel, Gerente de Distritos de Riego de la CONAGUA. Esta propuesta está lista para ser emitida a programas sectoriales como CONACYT-SEMARNAT o CONACYT-CONAGUA y poder ser presentada a las autoridades interesadas de CONAGUA o CFE en donde productos de pronóstico estacional les sean de potencial utilidad.

6. Bibliografía

Agresti, A., 2007. An Introduction to Categorical Data Analysis. Hoboken, New Jersey: JOHN WILEY AND SONS.

Brito-Castillo L, Leyva-Contreras A, Douglas AV, Lluch-Belda D., 2002. Pacific-decadal oscillation and the filled capacity of dams on the rivers of the Gulf of California continental watershed. Atmósfera, 15, 121–138.

Brockman, M.J., Wright, T.S., 1992. Statistical motor rating: making efficient use of your data. Journal of the Institute of Actuaries, 119 (3), 457-543.

Chambers, J.M., Trevor J., 1991. Hastie, editors. Statistical models in S. Chapman & Hall, 1st edition. ISBN 0-412-05291-1.

Delworth, T. y M. Mann, 2000. Observed and simulated multidecadal variability in the North Hemisphere. Climate Dynamics, 16, 661-676.

Dijkstra, H, L. Raa, M. Schmeits, J.Gerrits, 2006. On the physics of the Atlantic Multidecadal Oscillation. Ocean Dynamics, 56, 36-50.

Dobson, A.J., 2002. An Introduction to Generalized Linear Models. Boca Raton, Florida: CHAPMAN & HALL/CRC.

Enfield, D.B., A.M. Mestas-Nuñez, P.J. Trimble, 2001. The Atlantic multidecadal oscillation and its relation to rainfall and rivers flows in the continental U.S. Geophysical Research Letters, 28(10), 2077-2080.

Englehart PJ, Douglas AV., 2002. México's summer rainfall patterns: an analysis of regional modes and changes in their teleconnectivity. Atmósfera, 15, 147–164.

Faraway, J.J., 2006. Extending the Lineal Model with R. Boca Raton, Florida: CHAPMAN & HALL/CRC.

Guillén Estany, M., Ayuso Gutiérrez, M., Bolancé Losilla, C., Bermúdez Morata, L., Morillo López, I., Albarrán Lozano, I., 2005. El seguro de automóviles: estado actual y perspectiva de la técnica actuarial. Madrid: FUNDACIÓN MAPFRE ESTUDIOS.

Gochis, D.J. L. Brito-Castillo, W.J. Shuttleworth, 2007. Correlations between sea-surface temperatures and warm season streamflow in northwest Mexico. Int. J. Climatol., 27, 883–901.

Haberman, S., Renshaw, A.E., 1996. Generalized Linear Models and Actuarial Science. Journal of the Royal Statistical Society D, 45 (4), 407-436.

Hammer, G. L., Holzworth, D. P. & Stone, R. C. 1996. The value of skill in seasonal climate forecasting to wheat crop management in a region with high climatic variability. Aust. J. Agric. Res. 47, 717–737.

Hansen, J. W., Challinor, A., Ines, A. V. M., Wheeler, T., & Moron, V., 2006. Translating climate forecasts into agricultural terms: advances and challenges. Climate Research, 33(1), 27-41.

Hartmann D.L., 1994. Global Physical Climatology. Academic Press, 411 p.

Hastie T., Tibshirani, R., 1990. Generalized Additive Models. London: CHAPMAN AND HALL.

Hastie T., Tibshirani, R, Friedman, J., 2008. The elements of statistical learning. New York: SPRINGER.

Higgins RW, Shi W., 2001. Intercomparison of the principal modes of interannual and intraseasonal variability of the North American monsoon system. Journal of Climate 14: 403–417.

Hosmer, D.W.; Lemeshow, S., 2000. Applied Logistic Regression (2nd. Ed.). Hoboken, New Jersey: JOHN WILEY AND SONS.

Magaña, V., 2004 (ed.). Los Impactos de El Niño en México, 2ª ed., UNAM, 228p.

Magaña, V., J.L. Pérez y C. Conde, 1998. El fenómeno de El Niño y la Oscilación del Sur y sus impactos en México, Revista Ciencias 51:14-18.

Mantua, N.J., 2001. The Pacific Decadal Oscillation, Encyclopedia of Global Environmental Change, T. Munn (ed).

Mantua, N.J., S.R. Hare, Y. Zhang, J.M. Wallace y R.C. Francis, 1997. A Pacific decadal climate oscillation with impacts on salmon. Bulletin of the American Meteorological Society, 78:1069-1079.

McCullagh, P., Nelder, J.A., 1989. Generalized Linear Models. Boca Raton, Florida: CHAPMAN & HALL/CRC.

Minobe, S., 1997. A 50-70 year climatic oscillation over the North Pacific and North America. Geophysical Research Letters 24:683-686.

Minobe, S., 1999. Resonance in bidecadal and pentadecadal climate oscillations over the North Pacific: Role in climatic regime shifts. Geophysical Research Letters, 26:855-858.

Mo, K.C., Juang, H.H., 2003. Relationships between soil moisture and summer precipitation over the Great Plains and the Southwest. Journal of Geophysical Research 108: doi: 10.1029/2002JD002952. issn: 0148-0227.

Mosiño, P. y E. García, 1974. The Climate of Mexico, World Survey of Climatology, Vol. 11 Climates of North America, Bryson y Hare (eds), Elsevier, 345-404 p.

Mingfang, T., Kushnir, Y., Seager, Cuihua Li, C., 2009. Forced and Internal Twentieth-Century SST Trends in the North Atlantic. Journal of Climate, 22 (6), 1469–1481.

Nash, J. E., Sutcliffe, J. V., 1970. River flow forecasting through conceptual models part I — A discussion of principles, Journal of Hydrology, 10 (3), 282–290, DOI:10.1016/0022-1694(70)90255-6.

Nelder, J.A., Wedderburn, R.W.M., 1972. Generalized Linear Models. Journal of the Royal Statistical Society A, 135 (3), 370-384.

Ohlsson, E., Johansson, B., 2010. Non-Life Insurance Pricing with Generalized Linear Models. Heidelberg: SPRINGER.

Oliver, J. y J. Hidore, 2002. Climatology: An atmospheric science, USA: Prentice Hall, 410p.

Ojeda Bustamante, W., 2002. Importancia de la planeación hidroagrícola en distritos de riego bajo condiciones de baja disponibilidad. Revista AquaForum. Comisión Estatal de Agua de Guanajuato. Año 6 No. 30. pp. 10-14.

Philander, S.G., 1989. El Niño, La Niña, and the Southern Oscillation, USA: Academic Press, 293p.

Potgieter A.B, Hammer G.L, Butler D., 2002. Spatial and temporal patterns in Australian wheat yield and their relationship with ENSO. Aust. J. Agric. Res. 53, 77–89

Rebgetz, M. D., Chiew, F. H. S., & Malano, H. M. M., 2005. An investigation of the potential benefits of hydroclimate forecasts for irrigators in Northern Victoria. In MODSIM 2005 International Congress on Modelling and Simulation, Modelling and Simulation Society of Australia and New Zealand, Melbourne.

Santana Sepúlveda, J.S., Mateos Farfán, E., 2014. El arte de programar en R: un lenguaje para la estadística. Instituto Mexicano de Tecnología del Agua (IMTA), Jiutepec, México, 1ª edición. ISBN 978- 607-9368-15-9.

Schlesinger, M. E., 1994. An oscillation in the global climate system of period 65-70 years. Nature, 367 (6465), 723–726.

Stone, R.C, Hammer, G.L, Marcussen T., 1996. Prediction of global rainfall probabilities using phases of the Southern Oscillation Index. Nature, 384, 252–255.

Stone, R.C., Meinke, H., 2005. Operational seasonal forecasting of crop performance. Philosophical Transactions of the Royal Society of London B: Biological Sciences, 360, 2109-2124.

Trenberth, K.E., 1997. The Definition of El Niño, Bulletin of the American Meteorological Society, 78(12), 2771-2777.

Trenberth, K.E., D.P. Stepaniak , 2001. 'Indices of El Niño', Journal of Climate 14:1697-1701.

Vázquez-Aguirre, J.L. 2007. Variabilidad de la precipitación en la República Mexicana, Tesis de Maestría, Posgrado en Ciencias de la Tierra, Centro de Ciencias de la Atmósfera UNAM, 110pp.

Vittinghoff, E., Shiboski, S.C., Glidden, D.V., McCulloch, C.E., 2005. Regression Methods in Biostatistics. New York: SPRINGER.

Wang, Y., 2011. Smoothing Splines. Methods and Applications. Boca Raton, Florida: CHAPMAN & HALL/CRC.

Westra,S., Sharma, A., Brown, C., Lall, U., 2008. Multivariate streamflow forecasting using independent component analysis. Water Resour. Res., 44, W02437, doi:10.1029/2007WR006104.

Wood, S.N., 2006. Generalized Additive Models: An Introduction with R. Chapman & Hall/CRC, Boca Raton, Florida, first edition, 2006. ISBN 1-58488-474-6.

Yee, T.W., C. J. Wild, 1996. Vector Generalized Additive Models. Journal of Royal Statistical Society, Series B, 58(3), 481-493.

Yee, T.W., 2014. VGAM: Vector Generalized Linear and Additive Models. R package version 0.9‐5. URL http://CRAN.R‐project.org/package=VGAM.

7. Apéndice A. Código y salida del modelo estadístico de pronóstico de escorrentía con 2 meses de antelación

```
R version 3.1.1 (2014-07-10) -- "Sock it to Me"
Copyright (C) 2014 The R Foundation for Statistical Computing
Platform: x86_64-w64-mingw32/x64 (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

> ############## SEASONAL FORECASTS ####################
>
> #setwd("~/Dropbox/PROJECTS/MEXICO/Data")
> setwd("C:\\Mexico\\Data")
> dir()
 [1] "~$Results1.xlsx"                  "2014_Mexico_GLM_Forecast_index.R" "All_Data.csv"
 [4] "All_Data2.csv"                    "Challenge.docx"                   "Data_tool.R"
 [7] "Exp2a_flow2+rain2+ao2"            "Exp2n_flow2+rain2+nao2"
"Exp2na_flow2+rain2+ao2+nao2"
[10] "Exp3a_flow3+rain3+ao3"            "Exp3n_flow3+rain3+nao3"
"Exp3na_flow3+rain3+ao3+nao3"
[13] "Nino12.2+AO.2"                    "Nino4.2+AO.2"                     "Results1.xlsx"
> require(mgcv)
Loading required package: mgcv
Loading required package: nlme
This is mgcv 1.8-0. For overview type 'help("mgcv-package")'.
>
> data1 <- read.csv("All_Data2.csv")
>
> head(data1)
  Year Month   Flow Rain NINO12 NINO3 NINO4 NINO34      IOD   PDO    AMO  SWM   NAO     AO Flow.1
Rain.1 NINO12.1 NINO3.1
1 1959     4  17.04   NA  25.89 27.58 28.66  27.96 -0.561589 -0.02  0.014 1.39 -1.27  0.119  20.88
NA    27.07   26.79
2 1959     5   8.51   NA  24.26 26.94 29.05  27.92 -0.613850  0.23  0.024 0.63  0.42 -0.341  17.04
NA    25.89   27.58
3 1959     6  16.98   NA  22.63 26.03 28.39  27.37 -0.861683  0.44 -0.042 0.39  1.82 -0.033   8.51
NA    24.26   26.94
4 1959     7 130.40   NA  21.45 24.96 28.29  26.62 -1.151410 -0.50 -0.014 0.53  1.16  0.105  16.98
NA    22.63   26.03
5 1959     8 345.10   NA  20.34 24.32 28.29  26.33 -0.999596 -0.62  0.029 1.68  0.74 -0.745 130.40
NA    21.45   24.96
6 1959     9 141.60   NA  20.41 24.26 28.28  26.04 -0.890868 -0.85  0.143 0.82  0.77 -0.281 345.10
NA    20.34   24.32
  NINO4.1 NINO34.1     IOD.1 PDO.1  AMO.1 SWM.1 NAO.1   AO.1 Flow.2 Rain.2 NINO12.2 NINO3.2 NINO4.2
NINO34.2      IOD.2
1   28.44    27.24 -0.075042 -0.95 -0.010  0.71 -1.58  1.432  10.97     NA    25.68   26.43   28.50
27.30 -0.0437966
2   28.66    27.96 -0.561589 -0.02  0.014  1.39 -1.27  0.119  20.88     NA    27.07   26.79   28.44
27.24 -0.0750420
3   29.05    27.92 -0.613850  0.23  0.024  0.63  0.42 -0.341  17.04     NA    25.89   27.58   28.66
27.96 -0.5615890
4   28.39    27.37 -0.861683  0.44 -0.042  0.39  1.82 -0.033   8.51     NA    24.26   26.94   29.05
27.92 -0.6138500
```

```
5   28.29    26.62 -1.151410 -0.50 -0.014  0.53  1.16  0.105  16.98     NA    22.63   26.03   28.39
27.37 -0.8616830
6   28.29    26.33 -0.999596 -0.62  0.029  1.68  0.74 -0.745 130.40     NA    21.45   24.96   28.29
26.62 -1.1514100
  PDO.2  AMO.2 SWM.2 NAO.2   AO.2 Flow.3 Rain.3 NINO12.3 NINO3.3 NINO4.3 NINO34.3      IOD.3 PDO.3
AMO.3 SWM.3 NAO.3
1 -0.43  0.132  1.15 -1.00  2.544  16.02     NA    24.09   25.50   28.96    27.06 -0.2387870  0.69
0.118  0.66  0.37
2 -0.95 -0.010  0.71 -1.58  1.432  10.97     NA    25.68   26.43   28.50    27.30 -0.0437966 -0.43
0.132  1.15 -1.00
3 -0.02  0.014  1.39 -1.27  0.119  20.88     NA    27.07   26.79   28.44    27.24 -0.0750420 -0.95 -
0.010  0.71 -1.58
4  0.23  0.024  0.63  0.42 -0.341  17.04     NA    25.89   27.58   28.66    27.96 -0.5615890 -0.02
0.014  1.39 -1.27
5  0.44 -0.042  0.39  1.82 -0.033   8.51     NA    24.26   26.94   29.05    27.92 -0.6138500  0.23
0.024  0.63  0.42
6 -0.50 -0.014  0.53  1.16  0.105  16.98     NA    22.63   26.03   28.39    27.37 -0.8616830  0.44 -
0.042  0.39  1.82
    AO.3
1 -2.013
2  2.544
3  1.432
4  0.119
5 -0.341
6 -0.033
>
> ## select month you wish to forecast
> ## and select non-zero data
> data2 <- data1[data1$Flow>0&data1$Month ==9,]
> head(data2)
   Year Month  Flow  Rain NINO12 NINO3 NINO4 NINO34       IOD   PDO    AMO  SWM   NAO     AO Flow.1
Rain.1 NINO12.1 NINO3.1
6  1959     9 141.6    NA  20.41 24.26 28.28  26.04 -0.890868 -0.85  0.143 0.82  0.77 -0.281  345.1
NA    20.34   24.32
18 1960     9 107.6    NA  20.54 24.82 28.39  26.63 -1.278590 -0.94  0.228 0.99 -0.60 -0.382  154.8
NA    20.37   24.89
30 1961     9 281.9 120.3  19.75 23.61 28.27  25.93  3.071810 -2.01  0.020 0.46 -0.45  0.815  239.8
131.3    19.96   24.27
42 1962     9 291.9 131.6  20.16 24.21 28.14  26.10 -0.430731 -1.58  0.023 0.63 -1.99 -0.056  107.7
144.2    20.37   24.94
54 1963     9 124.6  51.3  21.15 25.44 28.82  27.42  1.732020  0.45 -0.183 1.13 -1.25  0.083  244.9
351.9    21.45   25.85
66 1964     9 352.2 148.8  19.51 23.94 27.13  25.52 -2.136670 -0.68 -0.201 1.31  1.54 -0.227  213.1
154.6    19.20   23.81
   NINO4.1 NINO34.1     IOD.1 PDO.1  AMO.1 SWM.1 NAO.1   AO.1 Flow.2 Rain.2 NINO12.2 NINO3.2 NINO4.2
NINO34.2     IOD.2
6    28.29    26.33 -0.999596 -0.62  0.029  1.68  0.74 -0.745 130.40     NA    21.45   24.96   28.29
26.62 -1.151410
18   28.18    26.90 -1.148280 -0.38  0.353  0.87  0.26 -1.008  56.81     NA    21.01   25.37   28.44
27.07 -1.043620
30   28.24    26.45  2.864480 -1.13  0.053  0.51 -1.97  0.013 306.00  282.1    20.55   24.82   28.45
26.90  2.170620
42   28.23    26.81 -0.606070 -0.48 -0.044  0.25 -1.66  0.152  90.85  130.1    20.57   25.31   28.38
27.05 -0.686428
54   28.81    27.62  1.298580 -1.03 -0.055  2.06  0.38 -0.625 435.60  360.5    22.01   26.43   29.01
28.01  0.668040
66   27.79    26.00 -2.205980 -1.03 -0.215  0.23 -0.58 -1.207  89.35  203.8    20.43   24.88   28.18
26.53 -1.902400
   PDO.2  AMO.2 SWM.2 NAO.2   AO.2 Flow.3 Rain.3 NINO12.3 NINO3.3 NINO4.3 NINO34.3     IOD.3 PDO.3
AMO.3 SWM.3 NAO.3
6  -0.50 -0.014  0.53  1.16  0.105  16.98     NA    22.63   26.03   28.39    27.37 -0.861683  0.44 -
0.042  0.39  1.82
18 -0.27  0.288  0.66  0.80 -0.619   9.29     NA    22.15   25.97   28.34    27.35 -1.221330  0.64
0.336  0.72  0.07
```

```
30 -1.22 -0.001  1.41 -1.47 -0.108  17.99  168.4    22.39   26.62   28.60    28.05  1.244080 -0.61
0.050  0.77 -0.96
42 -1.46  0.010  0.73 -0.85 -0.927  23.27   38.7    21.85   25.97   28.38    27.41 -0.627147 -1.62 -
0.053  1.34 -2.99
54 -1.00 -0.017  0.50 -1.04 -0.303   6.39   46.3    22.60   26.67   28.29    27.61  0.476605 -0.88 -
0.029  1.79 -0.43
66 -0.51 -0.125  0.89 -0.26  0.734  11.59   49.0    21.27   25.04   28.17    26.58 -1.442770 -0.32
0.022  1.27  1.39
     AO.3
6  -0.033
18  0.055
30  0.837
42  0.287
54 -0.585
66  0.142
>
> data2$bin <- ifelse(data2$Flow>median(data2$Flow)
+                     , 1, 0)
>
> ## Descriptive statistics
> ## normality
> plot(density(na.omit(data2$Flow))) ## or use a histogram
> hist(data2$Flow)
> ##summary
> summary(data2$Flow)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    7.0   124.6   167.1   199.4   244.4   597.1
>
> ## training
> data3 <- data2[1:(nrow(data2)/2),] ## First half of data
>
> # evaluate
> data4 <- data2[(nrow(data2)/2+1):(nrow(data2)+1),] ## Second half of data
>
> ## Continuous forecast ###############################################
>
> ## 2 month lag
> mod2 <- gam(Flow~s(Flow.2, k = 3)+s(Rain.2, k = 3)+s(NINO12.2, k = 3)
+             +s(AO.2, k = 3), data = data3, family = Gamma(link ="log"))
>
> summary(mod2)

Family: Gamma
Link function: log

Formula:
Flow ~ s(Flow.2, k = 3) + s(Rain.2, k = 3) + s(NINO12.2, k = 3) +
    s(AO.2, k = 3)

Parametric coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.28009    0.07058   74.81   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Approximate significance of smooth terms:
              edf Ref.df     F p-value
s(Flow.2)   1.471  1.719 0.597 0.53899
s(Rain.2)   1.000  1.000 0.011 0.91815
s(NINO12.2) 1.838  1.973 9.412 0.00215 **
s(AO.2)     1.900  1.989 6.964 0.00713 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

R-sq.(adj) =  0.451   Deviance explained = 72.9%
```

```
GCV = 0.17438  Scale est. = 0.10959   n = 22
> plot(mod2,pages = 1)
> gam.check(mod2)

Method: GCV   Optimizer: outer newton
full convergence after 10 iterations.
Gradient range [-1.471869e-07,1.067697e-08]
(score 0.1743807 & scale 0.1095865).
Hessian positive definite, eigenvalue range [1.47185e-07,0.00258846].
Model rank =  9 / 9

Basis dimension (k) checking results. Low p-value (k-index<1) may
indicate that k is too low, especially if edf is close to k'.

               k'   edf k-index p-value
s(Flow.2)   2.000 1.471   0.946    0.33
s(Rain.2)   2.000 1.000   1.105    0.64
s(NINO12.2) 2.000 1.838   0.888    0.22
s(AO.2)     2.000 1.900   1.469    0.98
>
> pred2 <- predict(mod2, newdata = data4, se.fit=T, type = "response")
>
> uci <- pred2$fit+2*pred2$se.fit
> lci <- pred2$fit-2*pred2$se.fit
>
> par(mfrow = c(2,1))
> plot(data4$Flow, pred2$fit, xlab="Observed",ylab="Predicted")
> plot(data4$Year, data4$Flow, xlab = "Year", ylab = "Flow Hm3", type = "l", ylim = c(0,1300))
> lines(data4$Year, pred2$fit, col = "red")
> lines(data4$Year, uci, col = "red", lty = 2)
> lines(data4$Year,lci, col = "red", lty = 2)
>
> ## Verify results of continuous forecast
> results <- na.omit(data.frame(obs = data4$Flow, pred = as.numeric(pred2$fit)))
>
> ## ME, MAE,...
>
> require(verification)
Loading required package: verification
Loading required package: fields
Loading required package: spam
Loading required package: grid
Spam version 0.41-0 (2014-02-26) is loaded.
Type 'help( Spam)' or 'demo( spam)' for a short introduction
and overview of this package.
Help for individual functions is also obtained by adding the
suffix '.spam' to the function name, e.g. 'help( chol.spam)'.

Attaching package: ‘spam’

The following objects are masked from ‘package:base’:

    backsolve, forwardsolve

Loading required package: maps
Loading required package: boot
Loading required package: CircStats
Loading required package: MASS
Loading required package: dtw
Loading required package: proxy

Attaching package: ‘proxy’

The following objects are masked from ‘package:stats’:
```

```
    as.dist, dist

Loaded dtw v1.17-1. See ?dtw for help, citation("dtw") for use in publication.

>
> verify(results$obs, results$pred, frcst.type = "cont", obs.type = "cont")
$baseline.tf
[1] FALSE

$MAE
[1] 64.60269

$MSE
[1] 7151.082

$ME
[1] 23.1394

$MSE.baseline
[1] 13974.63

$MSE.pers
[1] 36731.96

$SS.baseline
[1] 0.4882813

$obs
 [1] 182.53  90.30 185.68 135.44  91.45 247.02 214.18 167.04 232.85  72.80 324.01 147.54 136.93 597.12
99.24  97.86 167.74
[18]   7.00 187.59 155.35

$pred
 [1] 114.84250 183.66388 263.20762 169.07988 135.95920 320.39014 285.03595 220.28408 145.85315
122.07558 174.16392 154.10813
[13] 149.91289 517.45899 138.58565  92.17202 245.04173 252.43619 163.98643 154.20010

$baseline
[1] 176.9835

attr(,"class")
[1] "verify"    "cont.cont"
>
> ## NSE...
>
> require(hydroGOF)
Loading required package: hydroGOF
Loading required package: zoo

Attaching package: 'zoo'

The following objects are masked from 'package:base':

    as.Date, as.Date.numeric

Warning messages:
1: package 'hydroGOF' was built under R version 3.1.2
2: package 'zoo' was built under R version 3.1.2
>
> NSE(data4$Flow, as.numeric(pred2$fit))
[1] 0.1935414
>
> ## Binary forecast ###############################################
>
> ## only useful if there are months with and without
```

```
>
> modbin <- gam(bin~s(Flow.2, k = 3)
+               +s(Rain.2, k = 3)
+               +s(NINO12.2, k = 3)
+               +s(AO.2, k = 3)
+               , data = data3
+               , family = binomial)
> summary(modbin)

Family: binomial
Link function: logit

Formula:
bin ~ s(Flow.2, k = 3) + s(Rain.2, k = 3) + s(NINO12.2, k = 3) +
    s(AO.2, k = 3)

Parametric coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)    0.896      1.507   0.595    0.552

Approximate significance of smooth terms:
            edf Ref.df Chi.sq p-value
s(Flow.2)     1      1  2.995  0.0835 .
s(Rain.2)     1      1  3.941  0.0471 *
s(NINO12.2)   1      1  5.070  0.0243 *
s(AO.2)       1      1  2.603  0.1067
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

R-sq.(adj) =   0.76   Deviance explained =   79%
UBRE = -0.25622  Scale est. = 1         n = 22
> plot(modbin,pages = 1)
>
> ## Predict values
> pred4 <- predict(modbin, newdata = data4, type = "response")
> pred5 <- ifelse(pred4>0.5,1,0) ## predicted values
> obs <- data4$bin ## observed values
>
> ## Verify results of binomial forecast
> table.stats(obs, pred5, fudge = 0.01, silent = FALSE)
$tab

    0 1
  0 6 5
  1 3 6

$TS
[1] 0.4282655

$TS.se
[1] 0.1129117

$POD
[1] 0.6659267

$POD.se
[1] 0.1571346

$M
[1] 0.3329634

$F
[1] 0.4541326

$F.se
```

```
[1] 0.1500518

$FAR
[1] 0.4541326

$FAR.se
[1] 0.1003686

$HSS
[1] 0.2079105

$HSS.se
[1] 0.203805

$PSS
[1] 0.212904

$PSS.se
[1] 0.2172713

$KSS
[1] 0.2120998

$PC
[1] 0.5997001

$PC.se
[1] 0.1087322

$BIAS
[1] 1.222222

$ETS
[1] 0.1160221

$ETS.se
[1] 0.1269187

$theta
[1] 2.4

$log.theta
[1] 0.8754687

$LOR.se
[1] 0.9309493

$n.h
[1] 1.153846

$orss
[1] 0.4117647

$orss.se
[1] 0.3865534

$eds
[1] 0.3264547

$eds.se
[1] 0.2596339

$seds
[1] 0.1597809
```

```
$seds.se
[1] 0.2270099

$EDI
[1] 0.3200745

$EDI.se
[1] 0.528161

$SEDI
[1] 0.30155

$SEDI.se
[1] 0.5087791

>
>
> ################## MONTHLY FORECASTS ############################
>
> data5 <- data1[data1$Flow>0,]
>
> head(data5)
  Year Month   Flow Rain NINO12 NINO3 NINO4 NINO34       IOD   PDO    AMO  SWM   NAO     AO Flow.1
Rain.1 NINO12.1 NINO3.1
1 1959     4  17.04   NA  25.89 27.58 28.66  27.96 -0.561589 -0.02  0.014 1.39 -1.27  0.119  20.88
NA    27.07   26.79
2 1959     5   8.51   NA  24.26 26.94 29.05  27.92 -0.613850  0.23  0.024 0.63  0.42 -0.341  17.04
NA    25.89   27.58
3 1959     6  16.98   NA  22.63 26.03 28.39  27.37 -0.861683  0.44 -0.042 0.39  1.82 -0.033   8.51
NA    24.26   26.94
4 1959     7 130.40   NA  21.45 24.96 28.29  26.62 -1.151410 -0.50 -0.014 0.53  1.16  0.105  16.98
NA    22.63   26.03
5 1959     8 345.10   NA  20.34 24.32 28.29  26.33 -0.999596 -0.62  0.029 1.68  0.74 -0.745 130.40
NA    21.45   24.96
6 1959     9 141.60   NA  20.41 24.26 28.28  26.04 -0.890868 -0.85  0.143 0.82  0.77 -0.281 345.10
NA    20.34   24.32
  NINO4.1 NINO34.1     IOD.1 PDO.1  AMO.1 SWM.1 NAO.1   AO.1 Flow.2 Rain.2 NINO12.2 NINO3.2 NINO4.2
NINO34.2       IOD.2
1   28.44    27.24 -0.075042 -0.95 -0.010  0.71 -1.58  1.432  10.97     NA    25.68   26.43   28.50
27.30 -0.0437966
2   28.66    27.96 -0.561589 -0.02  0.014  1.39 -1.27  0.119  20.88     NA    27.07   26.79   28.44
27.24 -0.0750420
3   29.05    27.92 -0.613850  0.23  0.024  0.63  0.42 -0.341  17.04     NA    25.89   27.58   28.66
27.96 -0.5615890
4   28.39    27.37 -0.861683  0.44 -0.042  0.39  1.82 -0.033   8.51     NA    24.26   26.94   29.05
27.92 -0.6138500
5   28.29    26.62 -1.151410 -0.50 -0.014  0.53  1.16  0.105  16.98     NA    22.63   26.03   28.39
27.37 -0.8616830
6   28.29    26.33 -0.999596 -0.62  0.029  1.68  0.74 -0.745 130.40     NA    21.45   24.96   28.29
26.62 -1.1514100
  PDO.2  AMO.2 SWM.2 NAO.2   AO.2 Flow.3 Rain.3 NINO12.3 NINO3.3 NINO4.3 NINO34.3      IOD.3 PDO.3
AMO.3 SWM.3 NAO.3
1 -0.43  0.132  1.15 -1.00  2.544  16.02     NA    24.09   25.50   28.96    27.06 -0.2387870  0.69
0.118  0.66  0.37
2 -0.95 -0.010  0.71 -1.58  1.432  10.97     NA    25.68   26.43   28.50    27.30 -0.0437966 -0.43
0.132  1.15 -1.00
3 -0.02  0.014  1.39 -1.27  0.119  20.88     NA    27.07   26.79   28.44    27.24 -0.0750420 -0.95 -
0.010  0.71 -1.58
4  0.23  0.024  0.63  0.42 -0.341  17.04     NA    25.89   27.58   28.66    27.96 -0.5615890 -0.02
0.014  1.39 -1.27
5  0.44 -0.042  0.39  1.82 -0.033   8.51     NA    24.26   26.94   29.05    27.92 -0.6138500  0.23
0.024  0.63  0.42
6 -0.50 -0.014  0.53  1.16  0.105  16.98     NA    22.63   26.03   28.39    27.37 -0.8616830  0.44 -
0.042  0.39  1.82
    AO.3
```

```
1 -2.013
2  2.544
3  1.432
4  0.119
5 -0.341
6 -0.033
>
> data5$bin <- ifelse(data5$Flow>median(data5$Flow)
+                     , 1, 0)
>
> ## Descriptive statistics
> ## normality
> plot(density(na.omit(data5$Flow))) ## or use a histogram
> hist(data5$Flow)
> ##summary
> summary(data5$Flow)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  0.020   5.685  22.100  73.840 107.700 597.100
>
> ## training
> data6 <- data5[1:(nrow(data5)/2),] ## First half of data
>
> # evaluate
> data7 <- data5[(nrow(data5)/2+1):(nrow(data5)+1),] #
>
> mod3 <- gam(Flow~s(Flow.2, by = Month, k = 3)
+             +s(Rain.2, by = Month, k = 3)
+             +s(NINO12.2, by = Month, k = 3)
+             +s(AO.2, by = Month, k = 3), data = data6, family = Gamma(link ="log"))
>
> summary(mod3)

Family: Gamma
Link function: log

Formula:
Flow ~ s(Flow.2, by = Month, k = 3) + s(Rain.2, by = Month, k = 3) +
    s(NINO12.2, by = Month, k = 3) + s(AO.2, by = Month, k = 3)

Parametric coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   2.8557     0.2372   12.04   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Approximate significance of smooth terms:
                   edf Ref.df      F  p-value
s(Flow.2):Month   1.902  1.989  6.737 0.001487 **
s(Rain.2):Month   1.842  1.974  7.216 0.000998 ***
s(NINO12.2):Month 1.922  1.993 13.366  3.4e-06 ***
s(AO.2):Month     2.922  2.994  6.230 0.000448 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Rank: 10/13
R-sq.(adj) =  -0.0738   Deviance explained =   38%
GCV = 1.6461  Scale est. = 1.8371    n = 229
> plot(mod3,pages = 1)
> gam.check(mod3)

Method: GCV   Optimizer: outer newton
full convergence after 15 iterations.
Gradient range [-1.281495e-09,2.022e-07]
(score 1.646115 & scale 1.83709).
Hessian positive definite, eigenvalue range [0.0009361742,0.002041665].
```

```
Model rank =  10 / 13

Basis dimension (k) checking results. Low p-value (k-index<1) may
indicate that k is too low, especially if edf is close to k'.

                      k'   edf k-index p-value
s(Flow.2):Month    3.000 1.902   0.844    0.31
s(Rain.2):Month    3.000 1.842   0.866    0.54
s(NINO12.2):Month 3.000 1.922   0.846    0.36
s(AO.2):Month      3.000 2.922   0.951    0.92
>
> pred3 <- predict(mod3, newdata = data7, se.fit=T, type = "response")
>
> uci <- pred3$fit+2*pred3$se.fit
> lci <- pred3$fit-2*pred3$se.fit
>
> par(mfrow = c(2,1))
> plot(data7$Flow, pred3$fit, xlab="Observed",ylab="Predicted")
> plot(data7$Year, data7$Flow, xlab = "Year", ylab = "Flow Hm3", type = "l", ylim = c(0,1300))
> lines(pred3$fit, col = "red")
> lines(uci, col = "red", lty = 2)
> lines(lci, col = "red", lty = 2)
>
> verify(data7$Flow, pred3$fit, frcst.type = "cont", obs.type = "cont")
$baseline.tf
[1] FALSE

$MAE
[1] 102.8872

$MSE
[1] 30918.01

$ME
[1] 43.99679

$MSE.baseline
[1] 11688.44

$MSE.pers
[1] 14067.5

$SS.baseline
[1] -1.645178

$obs
  [1]  74.87 349.21 156.77 111.62  19.85  12.94  52.53  10.85   5.17   4.56   2.37   3.78 189.41 161.87
330.30 227.26  19.86
 [18]   9.69  10.83   3.71   2.42   1.45   1.09   0.72  57.76  42.57  68.38  27.75  52.74  65.34  33.53
52.82  77.29  13.22
 [35]   3.75   3.93  59.79 184.38 182.53  70.16  56.34  29.34  26.29  10.88   3.63   2.02   2.78 119.80
351.88 285.77  90.30
 [52]  39.45  16.43 161.63 209.40  28.30  11.57   6.52   3.93  18.01 114.64 284.14 185.68 233.60  21.84
2.35   1.52   1.16
 [69]   4.43 177.52 163.83 135.44 377.39  18.20  22.41  12.77   5.94   2.48   1.92 122.25 246.52  91.45
30.12  44.35  14.55
 [86]   6.30   4.30   4.16   3.06  13.04 340.04 465.74 247.02  38.96  25.62  54.48 414.37 214.18  34.29
37.40  16.77   9.35
[103]  22.86 559.51 220.36 167.04 156.06  19.89  63.20  13.62 413.07 257.89 232.85  54.70  30.72  99.13
18.66  19.00  11.10
[120]   0.17   3.73 130.40 133.70  72.80  13.21   0.51   0.31   0.43   0.48   0.02   5.68 189.95 172.49
324.01  28.20  15.11
[137]   0.34  13.90  58.56 201.89 147.54   8.50   3.55   4.58   4.56 120.37 169.74 136.93   8.62  12.25
122.09 130.92 597.12
```

```
[154]  12.14   6.29   0.13  68.43   8.04  66.59  89.41  99.24  14.54  37.99  34.21   7.56   3.35   1.11
57.46 142.36  97.86
[171]  47.71   0.19   6.11  49.17 182.27 125.65 167.74  12.42   3.31   1.00   1.00   1.00   2.00   1.00
6.00   5.00   9.00
[188]   7.00   8.00   4.00   3.00  12.31   8.48  96.56 226.19 187.59  19.63   2.45   0.20  38.36 144.17
155.35  19.10   3.49
[205]   0.81   7.58   4.97

$pred
       256        257        258        259        260        261        262        263        264
265        266
 39.084303 156.453228 171.168197 179.159929 142.473615 115.628070  20.521378  31.662070  35.540009
26.854538  11.133327
       267        268        269        270        271        272        273        274        275
276        277
 34.899585  69.000508 232.042721 192.110139 162.722558  40.017221  27.439899  21.988933  25.115789
28.240294  20.141078
       278        279        280        281        282        283        284        285        286
287        288
 23.269914  30.153359  74.062982  91.447216 421.532886 335.262677 431.795124 243.386749  20.799712
18.214510  10.738999
       289        290        291        292        293        294        295        296        297
298        299
  4.278069   6.319868   3.664391   5.163643   6.355181  77.455909  95.004793 495.215220 208.510662
21.406036  26.677071
       300        301        302        303        304        305        306        307        308
309        310
 35.492540  26.467711   6.521974  33.810511 125.487679 207.596213 117.134532  70.255110 498.252232
348.559000  21.605594
       311        312        313        314        315        316        317        318        319
320        324
 29.364972  12.638801  19.329291  28.206519  47.097261 105.226647 315.817880 251.343076  52.965824
52.972152  26.352368
       325        326        327        328        329        330        331        332        333
334        335
  5.487557  11.690442  49.648255  94.056872 208.977423 162.087782 170.006931 190.636002  21.523319
21.359781  29.217091
       337        338        340        341        342        343        345        346        347
348        349
  8.793880   3.658869  27.346857 125.495882 409.015233 118.510248 289.053463  22.591121  25.878484
26.414814  17.943310
       350        351        352        353        354        355        357        364        365
366        368
 22.752756  43.605342  70.518155 139.376639  99.313499  83.905718 169.721294  76.748138 185.794898
486.072387  26.694298
       369        370        371        375        376        377        378        379        380
381        382
170.053628  22.241722  31.094894  17.001141  54.820939 194.892848 202.772342  46.991417  60.673586
37.235796  22.900857
       388        389        390        391        392        393        395        396        397
398        399
 56.264178 147.929435 174.362670 151.312912 129.079096 300.474558  24.860077  40.737889  15.715611
6.639985   6.510510
       400        401        402        403        404        405        406        407        410
411        412
 10.950243  94.289202 272.559656 175.318474 294.428519 298.470600  22.061067  26.157824   9.523849
16.023529  19.482238
       413        414        415        416        420        423        424        425        426
427        428
117.802209 169.168872 310.101607  58.015029  30.158188  51.458843  97.885993 167.807825 730.536674
44.937589  96.130449
       429        435        436        437        438        439        447        448        449
450        451
714.504992  40.063804  82.192640 171.282817 148.732050 197.100038  36.354691 104.922881 252.653112
163.333649 263.776341
```

```
       453        454        457        459        460        461        462        463        464
465        466
193.963068  22.837278  13.974840   9.649690  10.878855  21.419361  98.371023  79.481413 103.727057
142.751775  18.883757
       467        469        472        473        474        475        479        482        483
484        485
 16.535062   6.666308   5.718737  50.531040 359.786877 277.318898  25.652595   9.334087  51.381904
83.699289 141.269068
       486        487        489        490        491        492        493        494        495
496        497
100.078765  63.633527 250.840941  21.521595  26.714945  27.338457  18.182996  17.931078  22.681692
52.439676 592.223126
       498        499        500        501        502        507        508        509        510
511        513
257.614005 612.745333 845.000856 702.016279  21.768309  10.881080  74.881193 305.312658 453.647901
45.182695 203.978172
       516        520        521        522        523        524        525        526        527
 29.689338  45.340720 123.099887 461.406189 146.242474 149.150897 111.190456  20.890317  24.379073

$baseline
[1] 77.23556

attr(,"class")
[1] "verify"    "cont.cont"
>
> ## NSE...
>
> NSE(data7$Flow, as.numeric(pred3$fit))
[1] -0.3793147
>
```